Strojno učenje svojstava materijala na podacima različite točnosti

Krčelić, Fran

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:460693

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-25**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

Fran Krčelić

Strojno učenje svojstava materijala na podatcima različite točnosti

Diplomski rad

Zagreb, 2024.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

INTEGRIRANI PREDDIPLOMSKI I DIPLOMSKI SVEUČILIŠNI STUDIJ
FIZIKA; SMJER ISTRAŽIVAČKI

Fran Krčelić

Diplomski rad

Strojno učenje svojstava materijala na podatcima različite točnosti

Voditelj diplomskog rada: dr. sc. Ivor Lončarić

Ocjena diplomskog rada: _____

Povjerenstvo: 1. _____

2. _____

3. _____

Datum polaganja: _____

Zagreb, 2024.

Prije svega, htio bih zahvaliti mentoru dr. sc. Ivoru Lončariću na njegovom strpljenju i entuzijazmu koji mi je prenio. Također, zahvalan sam na savjetima i vjeri koju je imao u mene.
Ponajviše, zahvalio bih se svojoj obitelji te prijateljima koji su me neprestano motivirali i pružali potporu tijekom cijelog studija.

Sažetak

Ovaj rad istražuje primjenu ekvivarijantnih graf neuralnih mreža s nasumičnim Fourierovim značajkama u predikciji energije molekula. Kombinirajući dva seta podataka različite točnosti, ANI-1x i ANI-1ccx, sastavljenih od energija organskih molekula baziranih na H, C, N, O elementima, model je usmjeren na poboljšanje prediktivne moći kroz dublje razumijevanje međuatomske interakcije. ANI-1x sadrži podatke generirane teorijom funkcionala gustoće za molekule izabrane aktivnim učenjem, omogućujući bolju generalizaciju s manjim brojem podataka, dok ANI-1ccx uključuje visoko precizne energije računane teorijom spregnutih grozdova. Treniranjem s različitim omjerima ovih setova pokazano je da povećanje količine ANI-1x podataka u trening setu rezultira boljom prediktivnom moći modela. Također, diskutirana je primjena ovoga pristupa na složenije sustave kao što su molekularni kristali. Ovakva metoda otvara nove mogućnosti za preciznije modeliranje u fizici, kemiji i znanosti o materijalima.

Ključne riječi: neuralne mreže, ekvivarijantne graf neuralne mreže, nasumične Fourierove značajke, molekularna energija

Multi-fidelity machine learning of material properties

Abstract

This thesis explores the application of equivariant graph neural networks with random Fourier features in predicting molecular energy. By combining two datasets of different accuracies, ANI-1x and ANI-1ccx, consisting of energies of organic molecules based on H, C, N, and O elements, the model aims to enhance predictive power through a deeper understanding of interatomic interactions. ANI-1x contains data generated with density functional theory for molecules selected through active learning, which allows for better generalization with fewer data points, while ANI-1ccx provides highly accurate energy values calculated using coupled-cluster methods. Training with various ratios of these datasets demonstrated that increasing the amount of ANI-1x data in the training set results in improved predictive power of the model. Additionally, the application of this approach to more complex systems, such as molecular crystals, was discussed. This method opens new possibilities for more accurate modeling in physics, chemistry and materials science.

Keywords: neural networks, equivariant graph neural networks, random Fourier features, molecular energy

Sadržaj

1 Uvod **1**

2 Metode **4**

2.1 Behler-Parrinello neuralna mreža 6

2.2 ANI-1 8

2.3 ANI-1x 12

2.4 ANI-1ccx 15

2.5 E(n) Ekvivarijantne graf neuralne mreže 17

2.6 Nasumične Fourierove značajke 21

2.7 Finalni model i pristup korišten u radu 22

3 Rezultati i diskusija **26**

3.1 Raspodjela podataka 26

3.2 Rezultati 27

3.3 Diskusija 30

4 Zaključak **32**

Dodatak **34**

Literatura **35**

1 Uvod

Fizika materijala je interdisciplinarna grana koja se proteže na velikom rasponu prostornih skala. Najosnovnija je od njih promatranje individualnih atoma te 3D struktura koje oni formiraju stvarajući kemijske veze. Precizan opis i razumijevanje, a konačno i manipuliranje materijala na razini atoma, centralni je problem fizike materijala.

Uobičajeno nas zanimaju energije te sile na pojedine atome kako bismo mogli modelirati sustav, stoga nam je korisna fizikalna veličina ploha potencijalne energije (PPE). Ona ovisi isključivo o položajima atoma, a nalazimo ju koristeći se kvantnom fizikom. Problem je dakle riješiti Schrödingerovu jednadžbu za elektrone u pitanju (korištenje PPE-a povlači Born–Oppenheimer aproksimaciju). Nažalost, egzaktno rješenje ovakvoga problema nije moguće dobiti čak ni s najsofisticiranijim računalima današnjice, stoga se već duže vrijeme koriste računalne simulacije na atomskim razinama kojima aproksimativno rješavamo problem; istaknule su se dvije glavne grane pristupa.

Prva od spomenutih metoda bazirana je na tzv. funkcionalima gustoće (DFT od eng. Density Functional Theory) [1]. Iako ova metoda daje iznimne rezultate, ograničena je skalom sustava. Njezina složenost raste s trećom potencijom broja elektrona u sustavu te je stoga primjenjiva na skalama do nekoliko stotina atoma. Usprkos tome, koristeći se DFT-om identificirani su novi spojevi i prije nepoznate strukture koje su kasnije i eksperimentalno realizirane.

S druge strane, ideja je parametrizirati međuatomske interakcije koristeći se generaliziranim potencijalom tzv. ”poljem sile”. Potencijal je opisan dobro poznatim jednostavnim članovima (harmonički doprinos, Coulombova interakcija, itd.) čije parametre onda prilagođavamo sustavu te ga zovemo ”empiričkim”. Očekivano, ovim postupkom žrtvujemo dio preciznosti, no zauzvrat dobivamo na brzini, a modelu složenost postaje proporcionalna drugoj potenciji broja atoma. Takve simulacije proširuju nam skalu sustava i do nekoliko milijuna atoma [2]. Iako su korisne za neke sustave, npr. opisivanje DNA molekule, polja sile nisu dovoljno kompleksni potencijali kojima bismo mogli zadovoljiti točnost opisivanja veće klase materijala.

Kao i u mnogim drugim područjima posljednjih godina, strojno učenje našlo primjenu je i u ovome. Općenito postoje tri glavna tipa strojnog učenja:

- Nadzirano učenje metoda je treniranja modela na predviđenom skupu poda-

taka, a ishod je poznat. Ovakvo učenje koristi se kao forma regresije. Modelu je zadatak raditi predikcije bazirane na danim podatcima koje se mjere pomoću funkcije gubitka (eng. loss function). Treniranje je završeno kada su odstupanja predikcija od stvarnih rezultata zadovoljavajuće mala.

- Nenadzirano učenje je ono u kojem model trenira na setu podataka, no ishod nije *a priori* poznat. Cilj je pronaći "skrivene" poveznice i uzorke u setu bez potrebe nekog specificiranog rezultata mreže.

- Učenje s potkrepljenjem metoda je u kojoj model interagira s okruženjem i dobiva povratnu informaciju u formi pozitivnog ili negativnog potkrepljenja. Ovakvo se učenje koristi u robotici i industriji videoigara. Cilj je naučiti mrežu da poduzima radnje koje će maksimizirati konačnu nagradu.

U ovom radu koristimo se prvim tipom strojnog učenja. Ideja je nalaženje PPE-a koristeći se neuralnim mrežama. Prvi ovakav pokušaj za velike sustave napravio je Smith 1999. [3]. Ideju su generalizirali Behler i Parinello 2007. na primjeru silicija [4]. Do danas smo proširili primjenu na mnoge druge sustave [5–7]. Štoviše, pokazano je da su duboke neuralne mreže sposobne imati točnost pristupa *ab initio* metoda, a pri tome biti brže čak do pet redova veličina [8]. Ipak ograničavajuć je faktor što kompleksnost prostora konfiguracija eksponencijalno raste s brojem različitih kemijskih simbola [9].

Proces je sličan onome s poljima sile jer također radimo neke pretpostavke kao što su lokalnost i glatkoća funkcije potencijala. Ključna razlika je u tome što nema nikakvih pretpostavki o obliku interakcije, tj. ovisnosti potencijala o nekoj veličini (npr. poziciji atoma) na već predefinirani način. Jednom kada je potencijal prilagođen podatcima, možemo predvidjeti energije i sile na velikim atomskim skalama bez potrebe novih informacija. Štoviše, poznavanjem PPE-a imamo pristup svim veličinama koje o njoj ovise.

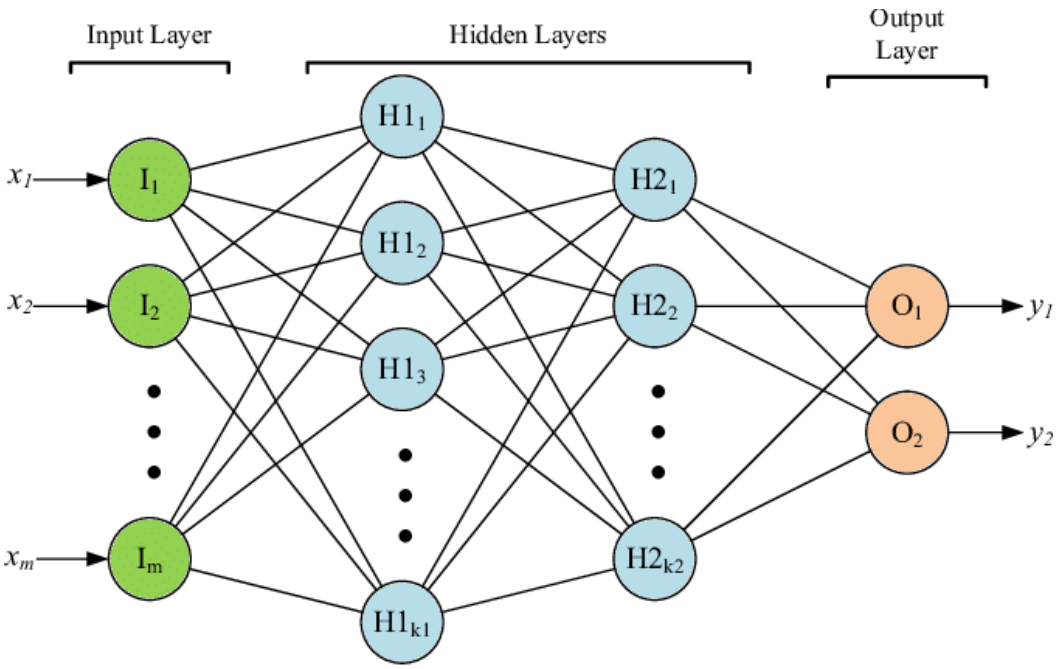
Nedostatak ovog pristupa dolazi od istog razloga kao i njegove prednosti. Time što prilagođavamo potencijal nekim podatcima imamo problem nepoznatog broja dovoljnih podataka da mreža daje zadovoljavajuće rezultate. Također, fizikalnost sustava nije *a priori* nametnuta, već sustav sam "uči" fiziku problema što može dovesti do besmislenih rezultata.

Zbog prethodno navedene kompleksnosti mreža s obzirom na broj različitih atoma, za potrebe ovog rada koristit ćemo samo četiri različita elementa (H, C, N, O). Kombinacijom ovih četiri elementa možemo konstruirati velik dio organskih spojeva koji će ovdje biti u fokusu. Naš je cilj istražiti metodu kombiniranja podataka različite točnosti u svrhu dobivanja boljeg modela u usporedbi s korištenjem samo manje količine ”boljih” podataka.

2 Metode

Prve ideje neuralnih mreža krenule su se javljati drugom polovicom prošlog stoljeća. One su tip strojnog učenja inspiriran načinom rada mozga. Bazična jedinica mreže je ”neuron” koji uzima informaciju, obrađuje je, te zatim šalje dalje (sljedećem neuronu). Doprinos pojedinog neurona određen je njegovim ”težinama” koje su promjenjivi numerički parametri. Pod izrazom ”treniranje mreže” podrazumijevamo modificiranje težina na način koji rezultira najboljim finalnim modelom. Taj uvjet je nametnut minimiziranjem funkcije srednjih kvadratnih pogrešaka (SKP) u odstupanju modela od već poznatih vrijednosti. Budući da je funkcija SKP direktno povezana s težinama, njihovim mijenjanjem prilagođavamo neuralnu mrežu i kažemo da ona u ovom slučaju ”uči” svojstva materijala.

Neuroni su grupirani u slojeve koji čine skup neurona koji međusobno ne komuniciraju, ali interagiraju sa susjednim slojevima. Na taj način izlaz jednog sloja neurona postaje ulaz za sljedeći. Slojeve dijelimo u tri skupine: ulazni, skriveni i izlazni slojevi (eng. input, hidden and output layer), shema je prikazana na slici 2.1. Ulazni sloj dobiva neobrađene podatke, dok izlazni sloj izbacuje konačni rezultat. Za razliku od njih, možemo imati više od jednog skrivenog sloja, što je većinom i slučaj. Oni su zaslužni za računalnu sposobnost mreže. Ukoliko mreža ima više od jednog skrivenog sloja, nazivamo ju dubokom neuralnom mrežom.



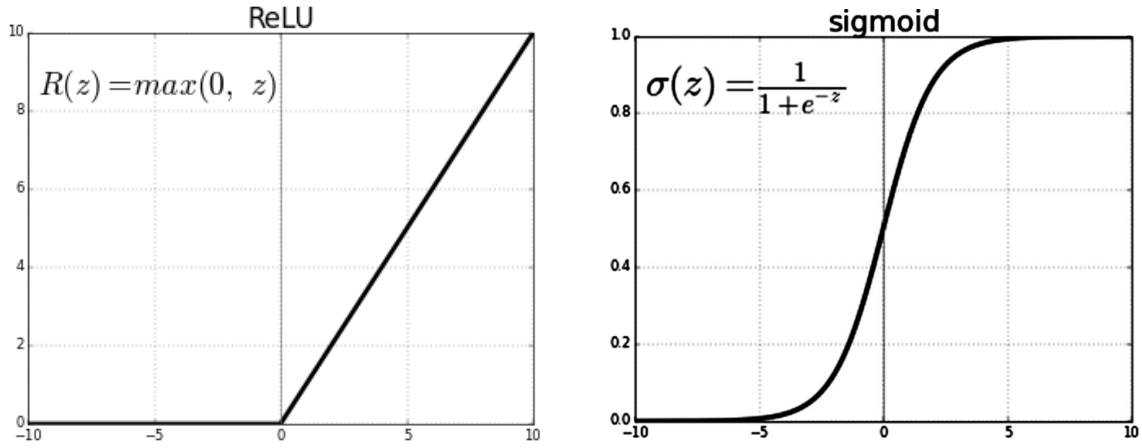
Slika 2.1: Shematski prikaz neuralne mreže.

Promotrimo što se ustvari događa pri obradi informacije u svakom od neurona. Već smo spomenuli da je izlaz prethodnog sloja ustvari ulaz sljedećeg, tj. neuron prima neki N-dimenzionalni vektor $\mathbf{x}$ gdje je N broj neurona prethodnog sloja. Prvi korak je njegova linearna transformacija

$$z = \mathbf{w} \cdot \mathbf{x} + b \tag{2.1}$$

gdje je $\mathbf{w}$ (također N-dimenzionalni) vektor prije spomenutih težina neurona, a b slobodni član. Kada bi ovo bio jedini korak, krajnji rezultat bio bi samo linearna transformacija ulaznog vektora. Budući da je kompozicija dvije linearne transformacije također linearna transformacija, nevezano koliko slojeva neuralne mreže iskoristimo, finalni rezultat uvijek će moći biti prikazan kao jedna operacija nad ulaznim parametrima. Time je neuralna mreža ograničena jer nema sposobnost opisivanja kompleksnijih ovisnosti, stoga je potreban drugi korak u obradi informacije u neuronu.

Nelinearnost dobivamo tako da izvrijednimo neku nelinearnu funkciju u točki z. Ovim postupkom omogućujemo neuralnoj mreži da opiše i složenije ovisnosti od linearne. Postoje razne nelinearne funkcije koje se koriste za ove potrebe, no zašto i kada koju koristimo nije predmet ovog rada. Više informacija na ovu temu može se naći u drugim izvorima [10]. Neke od najčešće korištenih nelinearnih funkcija su ReLU (eng. Rectified Linear Unit) (slika 2.2 lijevo) i sigmoid (slika 2.2 desno).



Slika 2.2: Aktivacijske funkcije.

Sada kada nam je jasan proces koji se odvija u pojedinom neuronu, promotrimo na koji način mijenjamo njihove težine, odnosno "treniramo" neuralnu mrežu. Kako se u ovom radu koristimo nadziranim učenjem, ulogu regresije imat će prethodno spomenuti pojam srednje kvadratne pogreške

$$\text{SKP} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \tag{2.2}$$

Veličina Y_i predstavlja pravu vrijednost varijable koju pokušavamo predvidjeti, $\hat{Y}_i$ predviđenu vrijednost našeg modela, a n broj primjera za koje radimo predikciju.

Još preostaje pitanje koje vrijednosti čine ulazni sloj neuralne mreže. Budući da se ovdje bavimo nadziranim učenjem, mreži ćemo dati uređeni par (X, Y), gdje je X ulazni vektor, a Y vektor pravih, već izračunatih vrijednosti. Specifičnije, u ovom slučaju X bi sadržavao pozicije atoma te njihove kemijske simbole, a Y bi predstavljao energiju te molekule izračunatu nekom drugom metodom, npr. DFT-om. Skup takvih uređenih parova nazivamo skupom za trening.

Naivno bismo mogli pomisliti da je format u kojem dajemo podatke neuralnoj mreži praktični neznačajan, dok god su sve informacije zapisane u njemu. Je li primjerice bolje koristiti Kartezijeve ili sferne koordinate ili neki drugi prikaz? Ispostavlja se da ni jedna od te dvije opcije nije pogodna. Ono što očekujemo je invarijantnost sustava na rotacije i translacije, te na zamjenu atoma istoimenih elemenata. Rješenje ovog problema nalazi se u korištenju grafova za prikaz molekula. Na taj način eliminiramo eksplicitne koordinate atoma u zamjenu za naglasak na veze među njima.

2.1 Behler-Parrinello neuralna mreža

Među prvima koji su riješili problem reprezentacije sustava bili su Behler i Parrinello u svom radu iz 2007. godine [4]. Ideja je bila transformirati Kartezijeve koordinate u simetrične funkcije (SF) G_i^{μ} koje će onda ispunjavati tražena svojstva. Index i je vezan uz index atoma, a μ poprima vrijednosti 1 za radijalni dio i 2 za angularni dio funkcije. Finalni izraz radijalnog dijela glasi

$$G_i^1 = \sum_{j\neq i}^{\text{atomi}} e^{-\eta\left(R_{ij}-R_s\right)^2} f_c\left(R_{ij}\right). \tag{2.3}$$

Parametar η predstavlja širinu Gausijana, a R_s vrijednost oko koje je on centriran. Funkcija f_c je ograničavajuća (eng. cutoff) funkcija koja ima ulogu da samo atomi u određenoj blizini promatranog atoma doprinose sumi. Njezin egzaktan izraz je

$$f_c(R_{ij}) = \begin{cases} \frac{1}{2}\cos\left(\frac{\pi R_{ij}}{R_C}\right) + \frac{1}{2}, & \text{za } R_{ij} < R_C \\ 0 & \text{za } R_{ij} > R_C. \end{cases} \tag{2.4}$$

Ova forma funkcije osigurava njezinu kontinuiranost te kontinuiranost njezine prve derivacije. Parametar R_C nazivamo ograničavajućim (eng. cutoff) radijusom; drugim riječima, samo atomi unutar sfere radijusa R_C doprinosit će sumi Gausijana. Ovakva transformacija koordinata osigurava da funkcije zadovoljavaju sva prethodno tražena svojstva simetrije jer ovisi samo o relativnoj udaljenosti atoma, a sadržava sve ključne informacije o interakciji između susjednih atoma.

Angularni dio simetričnih funkcija ima oblik

$$G_i^2 = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{atomi}} \left(1 + \lambda\cos\theta_{ijk}\right)^{\zeta} e^{-\eta\left(R_{ij}^2 + R_{ik}^2 + R_{jk}^2\right)} f_c(R_{ij}) f_c(R_{ik}). \tag{2.5}$$

Sada su u sumi prisutni tripleti atoma i, j i k te parametri ζ, λ i η koji nisu jedinstveni, odnosno njihove vrijednosti mogu biti korigirane po potrebi. Iznos kuta unutar kosinusa definiran je kao

$$\theta_{ijk} = \frac{\overrightarrow{R}_{ij} \cdot \overrightarrow{R}_{ik}}{R_{ij} R_{ik}}. \tag{2.6}$$

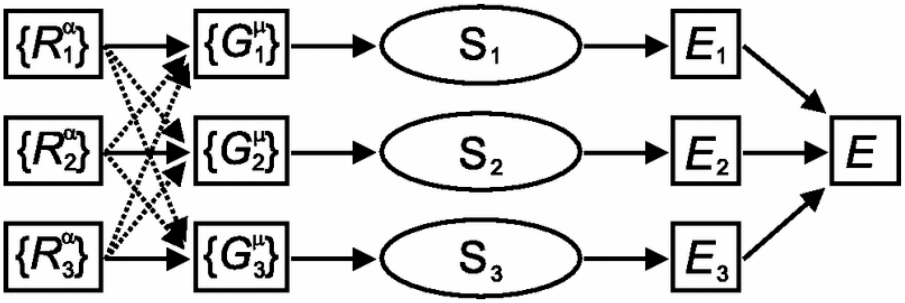
Kao i radijalna funkcija, i ova funkcija ovisi samo o relativnim udaljenostima atoma (R_{ij}) te time ispunjava jednaka simetrična svojstva. Naravno, ovo nije jedini mogući izbor funkcija koje zadovoljavaju tražena svojstva invarijantnosti te postoje brojne druge funkcije sa istim svojstvima.

Sada kada imamo reprezentaciju sustava preko simetričnih funkcija, sljedeći korak je izračunati energije pojedinih atoma.

Funkcije G_i^{μ} postaju ulazni parametri neuralnih mreža S_i, a izlazni parametri predstavljaju željene energije atoma E_i. Zbrajanjem energija pojedinih atoma dolazimo do ukupne energije sustava.

$$E = \sum_i E_i. \tag{2.7}$$

Shema rada mreže prikazana je na slici 2.3.



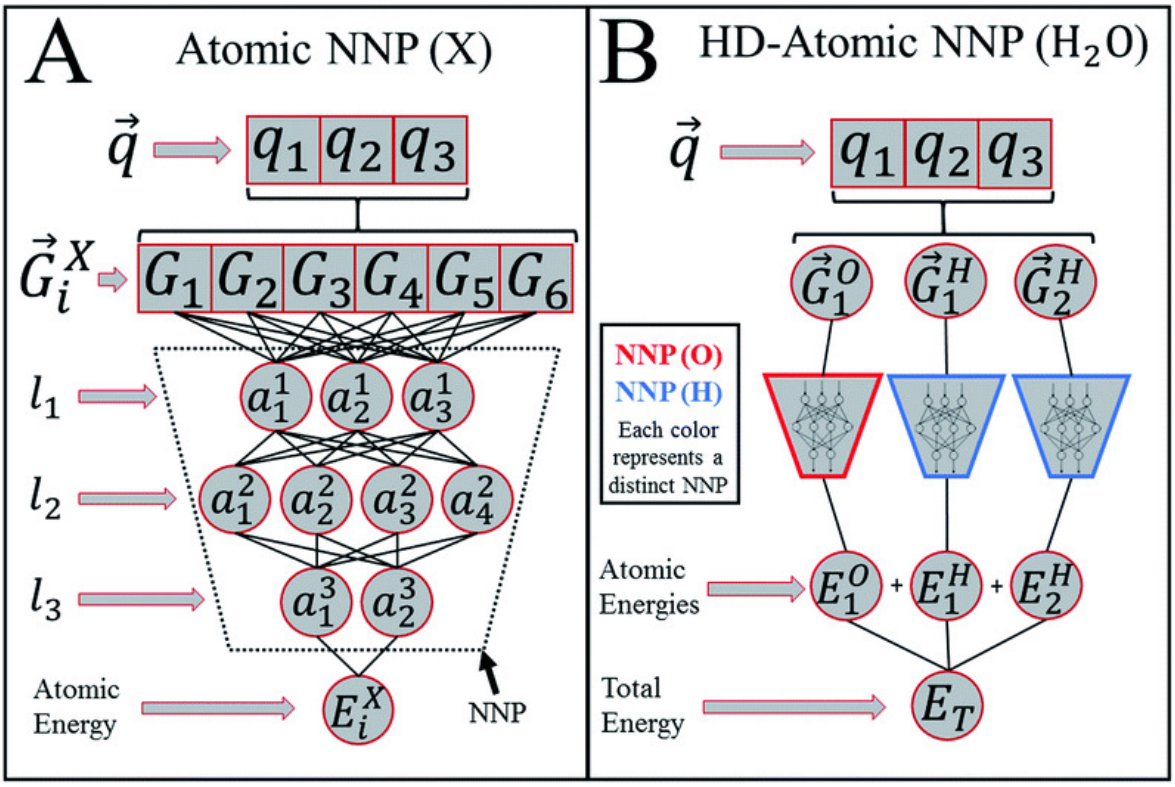
Slika 2.3: Shematski prikaz Behler-Parrinello neuralne mreže, preuzeto iz [4].

Bitno je napomenuti ukoliko su dva ili više atoma u sustavu istog kemijskog elementa, početne težine mreža S_i bit će postavljene na iste vrijednosti kako bi se očuvala invarijantnost na permutacije tih atoma.

2.2 ANI-1

Iako je Behler-Parrinello neuralna mreža imala usporedivu brzinu i čak bolju prediktivnu moć od postojećih polja sila, postojalo je još mnogo prostora za napredak. Jedan od većih problema bila je adaptivnost modela. Njihovi SF-ovi bili su primjenjivi samo na jedan kemijski sustav. Autori novog ANAKIN-ME (Accurate NeurAl networK engINe for Molecular Energies) modela [8] ili kraće ANI modela predstavili su dva glavna razloga za ograničenost BP modela. Prvo, SF-ovima nedostaje funkcionalni oblik za stvaranje prepoznatljivih značajki (prostorni raspored atoma koji se nalaze u uobičajenim organskim molekulama, npr. benzenski prsten, alkeni, funkcionalne skupine) u molekularnoj reprezentaciji, problem koji može spriječiti neuronsku mrežu od učenja interakcija u jednoj molekuli i zatim prijenosa tog znanja na drugu. Drugo, SF-ovi imaju ograničenu diferencijaciju atomskog broja, što empirijski otežava trening u složenim kemijskim okruženjima. Općenito, kombinacija ovih razloga ograničava izvorne SF-ove na kemijski simetrične sustave s jednom ili dvije vrste atoma ili vrlo male skupove podataka jedne molekule.

Autori ANI modela zamijenili su SF-ove s tzv. vektorima okoline atoma (VOA) koji su zadržali glavne ideje BP modela, no uz nekoliko modifikacija unaprijedili prethodni pristup. Shemu rada mreže možemo vidjeti na slici 2.4.



Slika 2.4: Shematski prikaz ANI-1 neuralne mreže, preuzeto iz [8].

Radijalni je dio oblikom ostao nepromijenjen

$$G_m^R = \sum_{j\neq i}^{\text{atomi}} e^{-\eta\left(R_{ij}-R_s\right)^2} f_c\left(R_{ij}\right) \tag{2.8}$$

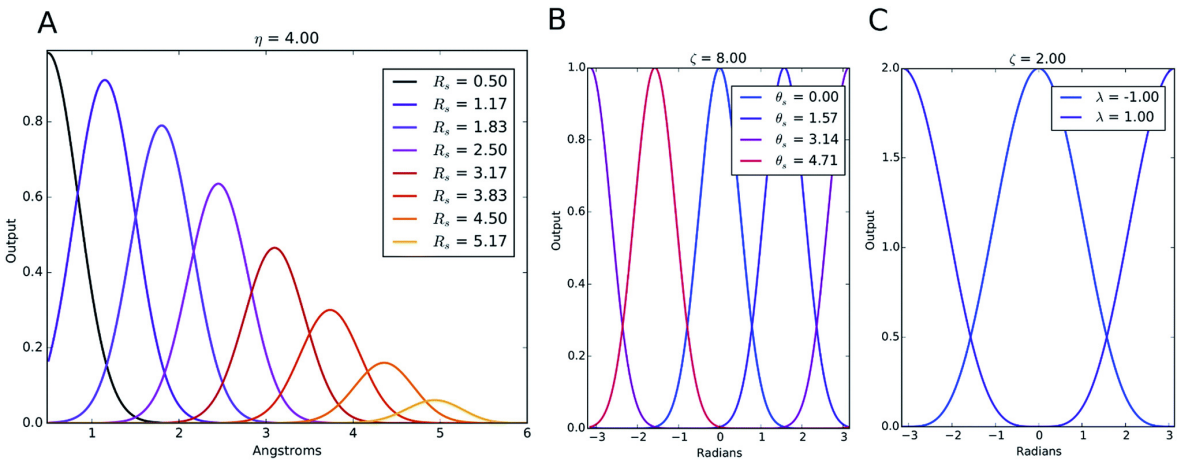
gdje je f_c definirana kao i u 2.4. Razlika je u indeksu m koji je ustvari uređeni par indeksa (η, R_s). Uloga parametara je identična kao i prije, η ima ulogu širine Gausijana, a R_s definira vrijednost oko koje je on centriran, no razlika je što sada ovi parametri nisu fiksni. Preciznije, R_s nije fiksan i mijenja se od atoma do atoma, a širina Gausijana ostavljena je fiksna zbog stabilnosti mreže, no ima mogućnost promjene.

Značajnije promjene nastale su u angularnom dijelu VOA

$$G_m^{A_{\text{mod}}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{atomi}} \left(1+\cos(\theta_{ijk}-\theta_s)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_c\left(R_{ij}\right) f_c\left(R_{ik}\right). \tag{2.9}$$

Sada indeks m predstavlja uređenu četvorku parametara $(\zeta, \theta_s, \eta, R_s)$. Novouvedeni parametar θ_s omogućava proizvoljan izbor faze, a R_s ima sličnu ulogu kao i kod

radijalnog dijela vektora. Iako se ove promjene ne čine drastičnima, one pridonose fleksibilnosti mreže, što joj omogućuje predviđanje energija i na molekulama koje su veće od onih u trening setu. Ovisnost simetričnih funkcija o odabranim parametrima prikazana je na slici 2.5.



Slika 2.5: Ovisnost SF-ova o različitim parametrima, preuzeto iz [8]. (A) Radijalni SF-ovi, (B) Modificirane angularni SF-ovi, (C) Originalne Behler-Parrinello angularni SF-ovi.

Sljedeći problem bio je izabrati povoljan set podataka za trening mreže. Autori su se odlučili za bazu podataka GDB-11 [11,12] koja sadrži sve molekule s do 11 atoma elemenata C, N, O i F. Za potrebe ANI-ja, odbačene su molekule koje sadrže fluor, te je broj "teških" atoma ograničen na 8. To je ostavilo 57 951 molekula pogodnih za treniranje mreže. Ovaj set nije bio dovoljno velik, no autori su vješto doskočili problemu. Budući da su sve molekule originalno dane u svojem osnovnom stanju, ideja je bila generirati nove primjere kroz pobuđivanje normalnih modova titranja. Od tuda dolazi i naziv uzorkovanje normalnih modova (NMS od eng. Normal Mode Sampling).

Prvi je korak ovog postupka izračunati normalne koordinate molekule. Ako molekula ima N_a atoma, onda su normalne koordinate

$$Q = \{q_1,\ q_2,\ q_3,\ \ldots,\ q_{N_f}\}. \tag{2.10}$$

Parametar N_f predstavlja broj modova molekule, koji iznosi $N_f = 3N_a - 5$ za linearne molekule, a $N_f = 3N_a - 6$ za sve ostale. Normalne koordinate računate su nekom od željenih *ab initio* metoda. Uz njih, računate su i pripadajuće konstante opruga

$$K = \{K_1,\, K_2,\, K_3,\, ...,\, K_{N_f}\}. \tag{2.11}$$

Još je preostalo nekako generirati pomake atoma iz ravnotežnog stanja. Za tu potrebu generiran je set pseudonasumičnih brojeva c_i tako da vrijedi

$$\sum_{i}^{N_f} c_i \in [0, 1] \tag{2.12}$$

te pomoću njih, generirani su pomaci za svaku normalnu koordinatu

$$R_i = \pm\sqrt{\frac{3c_i N_a k_b T}{K_i}} \tag{2.13}$$

tako da se harmonički potencijal podesi kao prosječna energija sustava čestica na nekoj temperaturi T, skaliran s c_i. Predznak pomaka određen je Bernoulijevom distribucijom gdje je $p = 0.5$ kako bi osigurali da su obje strane harmoničnog potencijala jednako zastupljene. Normalne koordinate skalirane su s dobivenim pomakom $q_i^R = R_i q_i$ te su nove konformacije generirane pomakom normalnih koordinata za pripadajuću vrijednost

$$Q_i^R = \{q_1 + q_1^R,\, q_2 + q_2^R,\, ...,\, q_{N_f} + q_{N_f}^R\}. \tag{2.14}$$

Zaključno, energije svih molekularnih konformacija izračunate su funkcionalom izmjene i korelacije ωB97x [13] sa 6-31G(d) kao baznim setom [14]. Iako postoje metode koje daju preciznije rezultate, njihova numerička cijena prevelika je za ovu količinu podataka. Time je originalni set podataka povećan praktički 300 puta te je sada na raspolaganju otprilike 17.2 milijuna konformacija.

Valja još spomenuti strukturu mreže koja je bila korištena te funkciju gubitka. Empirički, najbolja arhitektura mreže se pokazala kao $768 : 128 : 128 : 64 : 1$, tj. ona sa 768 ulaznih parametara, 3 skrivena sloja, te jednom vrijednosti u izlaznom sloju koji predstavlja energiju sustava. Najbolji ANI potencijal dobiven ovom mrežom nazvan je ANI-1. Pri treningu korištena je eksponencijalna funkcija gubitka

$$C(\overrightarrow{E}^{\text{ANI}}) = \tau \exp\left(\frac{1}{\tau}\sum_j \left(E_j^{\text{ANI}} - E_j^{\text{DFT}}\right)^2\right) \tag{2.15}$$

gdje je $\overrightarrow{E}^{\text{ANI}}$ vektor energija E_j^{ANI} koje model predviđa, a E_j^{DFT} pripadajuća ener-

gija izračunata DFT-om.

Iako je bio treniran na setu od samo 8 "teških" atoma, ANI-1 bio je primjenjiv i na sustave dvostruko veće od ovog. Ovako stvoreni potencijal bio je gotovo jednako precizan kao i DFT korišten za njegov trening i čak do 6 redova veličine brži u izračunu energija i sila na atome sustava. Također, empirički je pokazano da se numerička kompleksnost po atomu za velike molekule skalira ekvivalentno onome s poljem sila.

2.3 ANI-1x

Nakon ANI-1 modela, postavilo se pitanje kako dalje unaprijediti prediktivnu moć ove neuralne mreže. Očiti prijedlog bio bi povećanje kompleksnosti mreže, no uz to je usko vezana potrebna količina trening podataka što rezultira dužem vremenu treniranja. Ideja koja se pokazala uspješnom ustvari je bila smanjiti trening set, ali na smion način. Koristeći se tzv. aktivnim učenjem (AU), autori ANI-1 modela postižu bolje performanse sa samo 25% originalnog seta podataka, a novi model nazivaju ANI-1x [15]. Trenutni set trening podataka sadrži mnogo konformacija koje ne doprinose nikakvim novim informacijama mreži te samo usporavaju njezin proces treniranja. Ideja AU je a priori odabrati primjere od kojih će se sastojati trening set. Ovaj postupak odvija se u dvije glavne etape.

Prvi je korak odvojiti 2% podataka originalnog ANI-1 seta te trenirati model na njima. Preostali dio molekula je testiran njime te je 2% odbačenih primjera pripojeno novom trening setu. Molekulu smatramo odbačenom ukoliko model jako griješi pri predikciji njezine energije, a egzaktan kriterij dan je uvjetom

$$\left|E_{\text{ANI}} - E_{\text{DFT}}\right| / \sqrt{N} > 0.04\,\text{kcal/mol} \tag{2.16}$$

gdje je N broj atoma promatrane molekule. Ovaj postupak se iterativno ponavlja dok god postoji više od 5% primjera u testnom setu koji zadovoljavaju 2.16. U trenutku kada to prestane biti slučaj, preostalih $< 5\%$ primjera manualno se dodaju trening setu. Svi su korišteni parametri proizvoljni, te se ovaj proces može obaviti u finijim ciklusima uz cijenu vremena trajanja.

U drugom koraku počinje istraživanje konfiguracijskog prostora s pomoću nasumičnog uzorkovanja na primjerima koje model nije vidio. Za ovu potrebu, autori su

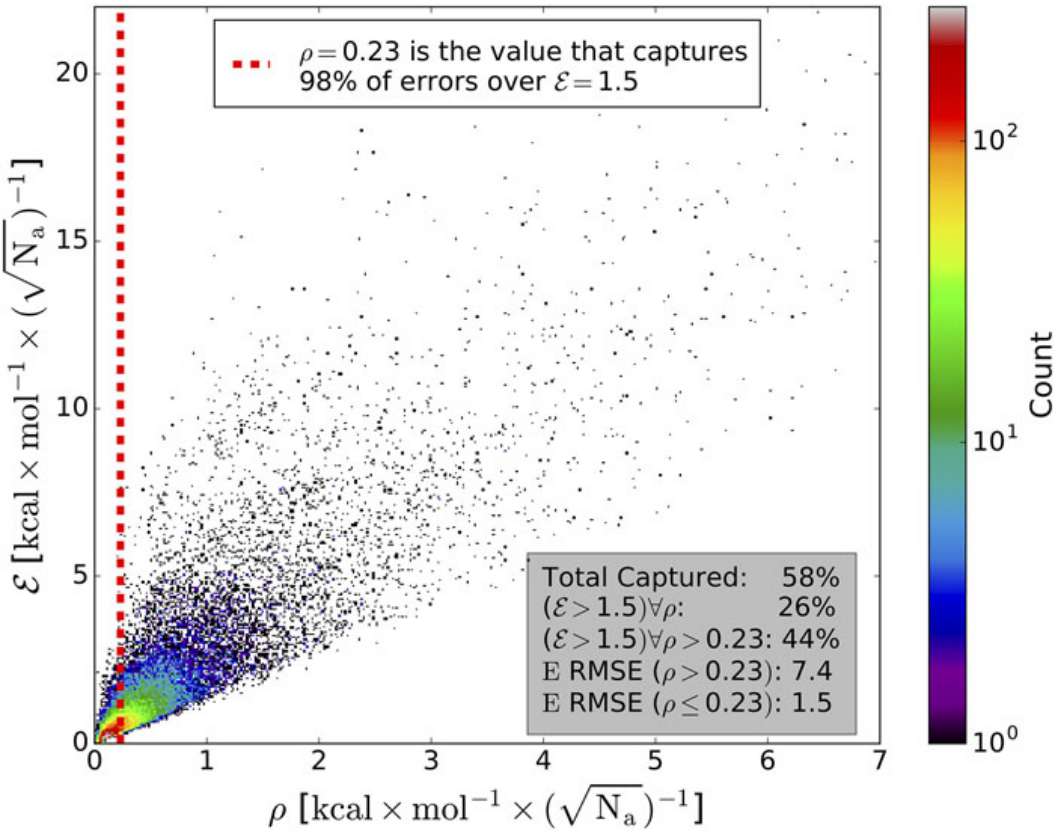
se koristili setovima podataka GDB-11 [11, 12] i ChEMBL [16–18]. Zatim je korišten ansambl od pet ANI-1 modela te je dobivena energija za svaki od njih. Ako se modeli slažu to je indikacija da je model sposoban opisati takav sustav i on nam nije od velikog značaja. S druge strane, ako modeli ne daju slične rezultate, dodavanje takve molekule u trening set bilo bi korisno modelu. Proces kojim se ovo odvija zove se upit od strane odbora (QBC od eng. Query by Committee) [19]. Test koliko se modeli u ansamblu međusobno slažu bit će standardna devijacija distribucije predikcija normalizirana s brojem atoma u molekuli

$$\rho_i = \frac{\sigma_i}{\sqrt{N_i}}. \tag{2.17}$$

U slučaju kada je ova vrijednost veća od neke predefinirane vrijednosti $\hat{\rho}$ molekula biva uključena u trening set. Za potrebe odabira $\hat{\rho}$ definirana je veličina

$$\varepsilon_i = \left| MAX\left(\left\{ E_T^{\text{ANI}} \right\}_i^{\text{ens}} - E_{T,i}^{\text{REF}} \right) \right| / \sqrt{N_i} \tag{2.18}$$

koja predstavlja maksimalno odstupanje modela ansambla od stvarne vrijednosti energije. Predikcije modela prikazane su na slici 2.6.



Slika 2.6: Empiričko određivanje granične vrijednosti, preuzeto iz [15].

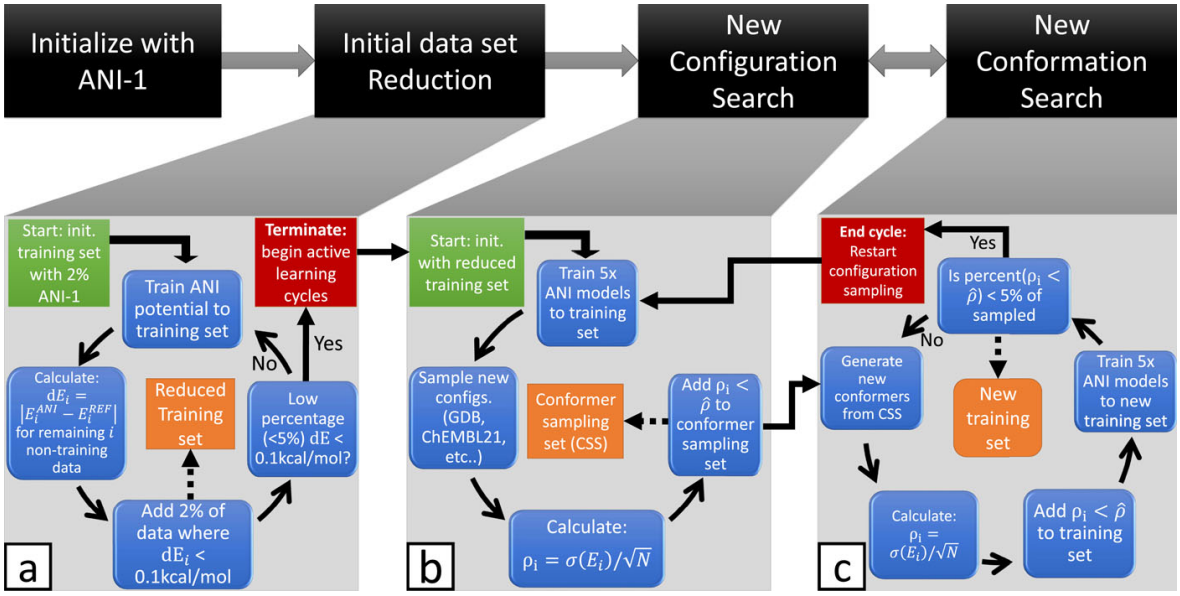
Postavlja se pitanje koje je maksimalno odstupanje modela koje će se tolerirati? Uzeta vrijednost bila je

$$\varepsilon_{\text{max}} = 1.5\,\text{kcal/mol}. \tag{2.19}$$

Zatim se postavlja pitanje koliki postotak molekula će zadovoljavati 2.19 za dani $\hat{\rho}$? Utvrđeno je da za vrijednost $\hat{\rho} = 0.23$ obuhvaćamo 98% primjera za koje vrijedi 2.19 te je ona uzeta kao granična. Ovim odabirom granične vrijednosti, 58% primjera palo je iznad dopuštene vrijednosti te su korišteni u novom trening setu.

Posljednji je korak optimizirati ove odabrane sustave u njihovo osnovno stanje te ponovno njihovom perturbacijom proširiti trening set. Korištene metode su: uzorkovanje različitih normalnih modova (eng. diverse normal mode sampling), uzorkovanje nasumičnih putanja (eng. random trajectory sampling) te uzorkovanje dimera generiranih molekulskom dinamikom (eng. molecular dynamics generated dimer sampling). Više o njima može se pročitati u izvornom članku [15]. Za izračun energija ovih konformacija ponovo je korišten funkcional ωB97x [13] sa 6-31G(d) kao baznim setom [14].

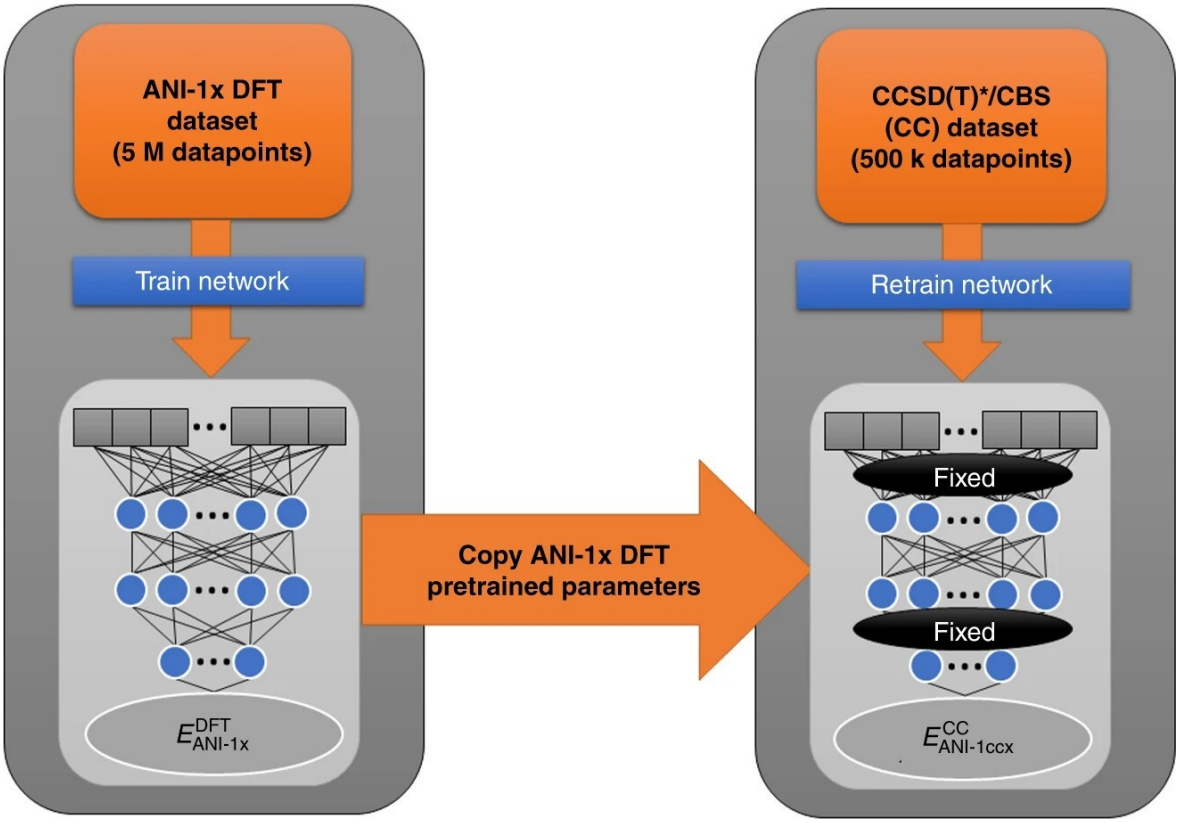
Ovaj ciklus se ponavlja dok god se poboljšava prediktivna moć modela. Finalna verzija dobivena ovom metodom nazvana je ANI-1x. Shema procesa prikazana je na slici 2.7. Iako koristi samo četvrtinu količine podataka u usporedbi s ANI-1, mreža pokazuje značajno poboljšanje prediktivne moći od svojeg prethodnika.



Slika 2.7: Shematski prikaz ANI-1x neuralne mreže, preuzeto iz [15].

2.4 ANI-1ccx

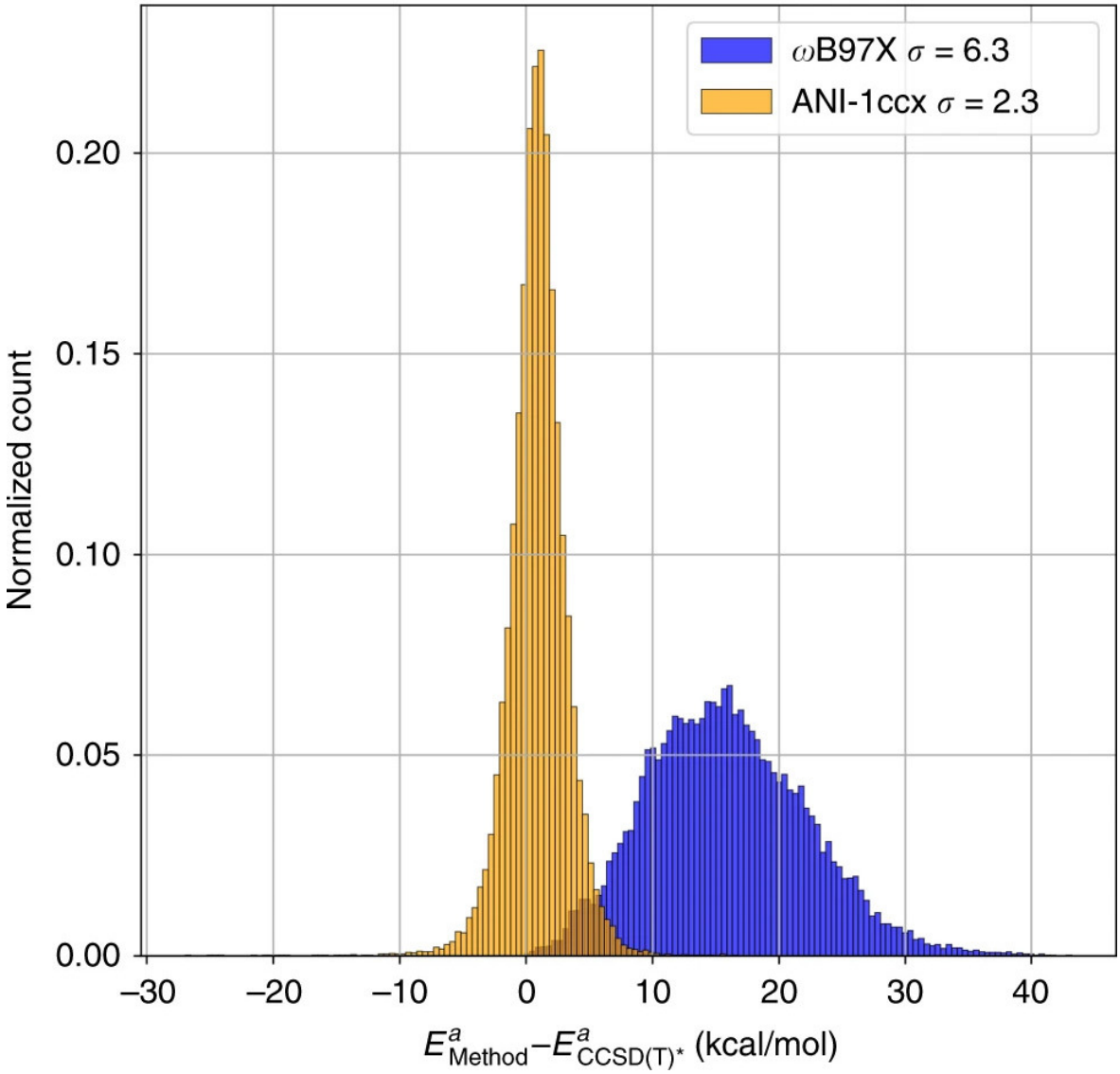
Teorija spregnutih grozdova (CC od eng. coupled clusters) sistematično pristupa egzaktnom rješenju Schrodingerove jednadžbe te se smatra zlatnim standardom za mnoge aplikacije [20–22]. Korištenjem spregnutih grozdova koji uzimaju u obzir jednostruke, dvostruke i perturbativne trostruke pobude (CCSD(T) od eng. coupled clusters considering single, double, and perturbative triple excitations) te ekstrapolacijom na cijeli bazni set (CBS od eng. complete basis set) [23, 24], moguće je opisati i najkompleksnije interakcije atoma u molekulama [25]. Nažalost, računanje na CCSD(T)/CBS razini računski je skupo i često nepraktično za sustave s više od desetak atoma. Ideja je razviti model koji bi mogao parirati preciznosti CC metode, a koji bi u isto vrijeme bio široko primjenljiv na druge sustave. Ovi su zahtjevi postignuti koristeći se prijenosnim učenjem (eng. transfer learning) te je novi model nazvan ANI-1ccx [26]. Ideja je prvo trenirati model na većem setu podataka te ga zatim finijim promjenama prilagoditi za traženi set podataka. Shema modela nalazi se na slici 2.8.



Slika 2.8: Shematski prikaz ANI-1ccx neuralne mreže, preuzeto iz [26].

Prvi korak bio je trenirati model po uzoru na ANI1-x, koristeći prije dobivenih 5 milijuna primjera za učenje te ga zatim doraditi na manjem setu od pola milijuna podataka s preciznije izračunatom energijom. Koristeći odabrane aproksimacije, izračunali su energije molekula s usporedivom preciznošću CCSD(T)/CBS metode, ali uz manju računsku cijenu. Više o tom postupku može se naći u izvornom radu [26].

Sada kada su sve komponente za prijenosno učenje prisutne, preostalo je još "dotrenirati" postojeći ANI-1x model na novim primjerima. Dva skrivena sloja mreže držana su fiksnima, a druga dva bila su otvorena promjeni. Preciznije, od 325 248 slobodnih parametara inicijalnog modela, 65 280 njih postavljeno je kao fiksno kako bi se izbjeglo pretreniranje mreže.



Slika 2.9: Usporedba predviđanja energije dvaju modela, preuzeto iz [26].

Na slici 2.9 prikazana je usporedba performanse ANI1-ccx modela s DFT-om (korištenim za potrebe ANI-1 i ANI-1x modela) nad konformacijama GDB-10 do 13, s do 100 kcal/mol odstupanjima od energetskog minimuma. Distribucija odstupanja energija za ANI1-ccx ima standardnu devijaciju od 2.3 kcal/mol, dok DFT ima devijaciju od 6.3 kcal/mol. Također, vidimo da je raspodjela ANI1-ccx centrirana simetrično oko nule. DFT s druge strane ima tendenciju premašiti realnu energiju sustava u prosjeku za 15-ak kcal/mol. Uz energije, ANI1-ccx korišten je za računanje sila, energija reakcije i izomerizacije te molekularne torzije, no to već izlazi izvan okvira našega rada.

2.5 E(n) Ekvivarijantne graf neuralne mreže

Kao što i sam naziv kaže, pristup ovih neuralnih mreža baziran je na ekvivarijantnosti, preciznije na ekvivarijantnosti rotacija, translacija, refleksija i permutacija grafova, što je upravo slučaj s primjerima molekula. Naziv modela potječe od grupe simetrija translacije i rotacije 3D tijela, koja se u teoriji grupa označava kao SE(3). Dodamo li toj grupi i refleksije, oznaka postaje E(3). Budući da se ovaj model može primijeniti i u više dimenzija, dana mu je oznaka E(n). Implementaciju ovakvog modela, napravili su Victor Garcia Satorras, Emiel Hoogeboom i Max Welling 2021. godine [27]. Ovaj model je bio jednostavniji od drugih sa sličnim pristupom jer je izbjegao korištenje sfernih harmonika. Također, kao što je već spomenuto, nije bio ograničen na samo trodimenzionalni prostor, a skaliranje kompleksnosti s dimenzijom nije bilo drastično.

Definirajmo za početak pojam ekvivarijantnosti malo preciznije. Neka je $T_g : X \to X$ set transformacija za apstraktnu grupu $g \in G$. Kažemo da je funkcija $\phi : X \to Y$ ekvivarijantna na g ako postoji ekvivalentna transformacija na kodomeni $S_g : Y \to Y$ takva da

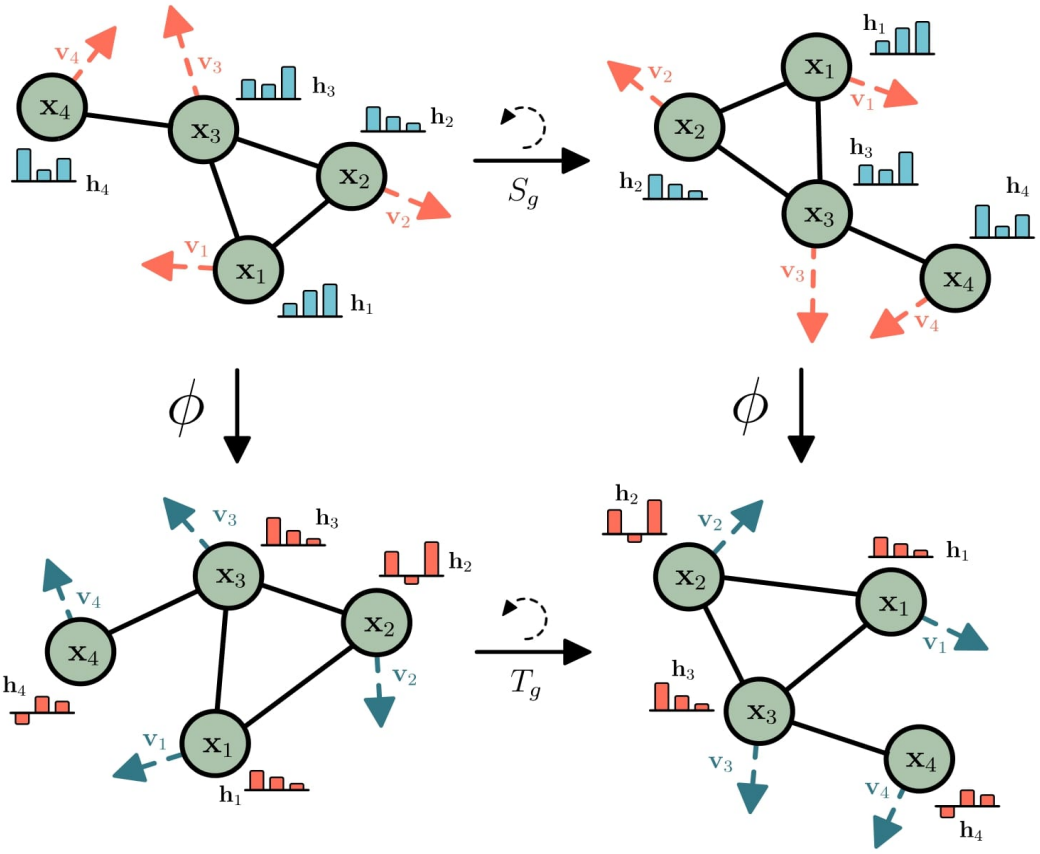
$$\phi\big(T_g(x)\big) = S_g\big(\phi(x)\big). \tag{2.20}$$

Uzmimo da je **x** neki n-dimenzionalni vektor, za tri prije spomenuta svojstva vrijede relacije

- Translacija **x** za neki $g \in \mathbb{R}^n$ rezultirat će istom translacijom u kodomeni, $\phi(\mathbf{x}) + g = \phi(\mathbf{x} + g)$.

- Rotacija (i refleksija) za neku matricu $Q \in \mathbb{R}^{n \times n}$. Vrijedit će $Q(\phi(\mathbf{x})) = \phi(Q\mathbf{x})$. Shematski prikazano na slici 2.10.

- Permutacija vektora domene pa preslikavanje daje isti vektor koji bi dobili da prvo preslikamo $\mathbf{x}$ pa primijenimo identičnu permutaciju na kodomeni, $P(\phi(\mathbf{x})) = \phi(P(x))$.



Slika 2.10: Shematski prikaz ekvivarijantnosti rotacije, preuzeto iz [27].

Tipične neuralne mreže na bazi grafova (GNN od eng. Graph Neural Network) rade na sljedećem principu. Neka je $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ graf s čvorovima $v_i \in \mathcal{V}$ i bridovima $\varepsilon_{ij} \in \mathcal{E}$, nad njim možemo definirati sloj konvolucije

$$\mathbf{m}_{ij} = \phi_e\big(\mathbf{h}_i^l, \mathbf{h}_j^l, a_{ij}\big) \tag{2.21}$$

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij} \tag{2.22}$$

$$\mathbf{h}_i^{l+1} = \phi_h\big(\mathbf{h}_i^l, \mathbf{m}_i\big) \tag{2.23}$$

gdje je $\mathbf{h}_i^l \in \mathbb{R}^k$ k-dimenzionalna reprezentacija čvora v_i u sloju l. Parametri a_{ij} su atributi bridova, a $\mathcal{N}_i$ predstavlja skup susjeda čvora v_i. Konačno, ϕ_e i ϕ_h su operacije na bridovima i čvorovima, koje se obično aproksimiraju s višeslojnim perceptronima (MLP od eng. Multilayer Perceptron).

Promotrimo sada kako se ekvivarijantne graf neuralne mreže (EGNN od eng. Equivariant Graph Neural Network) razlikuje od GNN-a. Koristeći se istom notacijom, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ je graf s čvorovima $v_i \in \mathcal{V}$ i bridovima $\varepsilon_{ij} \in \mathcal{E}$. Uz reprezentacije čvorova $\mathbf{h}_i \in \mathbb{R}^k$, sada su bitne i n-dimenzionalne koordinate $\mathbf{x}_i \in \mathbb{R}^n$ povezane sa svakim čvorom grafa. Uzmimo da graf ima M čvorova, tj $i \in \{1, 2, ..., M\}$. Model će očuvati ekvivarijantnost prema rotacijama i translacijama na ovim skupovima koordinata $\mathbf{x}_i$ te će također biti očuvana i ekvivarijantnost prema permutacijama na skupu čvorova $\mathcal{V}$, na isti način kao i kod GNN-a. Sloj EGNN-a definiran je jednadžbama

$$\mathbf{m}_{ij} = \phi_e\Big(\mathbf{h}_i^l, \mathbf{h}_j^l, \left|\left|\mathbf{x}_i^l - \mathbf{x}_j^l\right|\right|^2, a_{ij}\Big) \tag{2.24}$$

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + C \sum_{j \neq i} \big(\mathbf{x}_i^l - \mathbf{x}_j^l\big)\phi_x(\mathbf{m}_{ij}) \tag{2.25}$$

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij} \tag{2.26}$$

$$\mathbf{h}_i^{l+1} = \phi_h\big(\mathbf{h}_i^l, \mathbf{m}_i\big). \tag{2.27}$$

Operacija nad bridovima modificirana je dodavanjem relativne kvadratne udaljenost dviju koordinata $\left|\left|\mathbf{x}_i^l - \mathbf{x}_j^l\right|\right|^2$ te je dobivena jednadžba 2.24. Ulazni parametri; reprezentacije $\mathbf{h}_i^l$, $\mathbf{h}j^l$ i atributi bridova aij, ostaju isti. Za potrebe ovog rada, atributi bridova uključivat će vrijednosti bridova $a_{ij} = \varepsilon_{ij}$, ali mogu uključivati i dodatne informacije.

Izraz 2.25 predstavlja ažuriranje vektora pozicije svakog čvora. Kao što vidimo, pozicija se ažurira težinskim zbrojem svih relativnih udaljenosti $(\mathbf{x}_i - \mathbf{x}_j)$ za svaki j. Težine su odrađene funkcijom $\phi_x : \mathbb{R}^k \to \mathbb{R}$ koja kao ulazni parametar uzima reprezentaciju brida $\mathbf{m}_{ij}$ iz prethodne operacije i kao izlaz daje skalar. Konstanta C odabrana je kao $1/(M-1)$, što dijeli zbroj s brojem njegovih elemenata. Ovaj izraz glavna je razlika između EGNN-a i GNN-a te je razlog zašto su ekvivarijantnosti translacije i rotacije očuvane (dokaz se nalazi u izvornom radu [27]). Unatoč svojoj jednostavnosti, ova ekvivarijantna operacija je vrlo fleksibilna jer sada reprezentacija

$\mathbf{m}_{ij}$ može nositi informacije iz cijelog grafa, a ne samo iz zadanog brida ε_{ij}. Upravo zbog tih invarijantnosti, EGNN modeli se pokazuju bolji u modeliranju trodimenzionalnih problema u odnosu na ostale neuralne mreže zasnovane na temelju grafova.

Jednadžbe 2.26 i 2.27 slijede ista ažuriranja kao i standardni GNN-ovi. Jednadžba 2.26 grupira sve dolazne informacije od susjednih čvorova $\mathcal{N}(i)$ s obzirom na čvor v_i, a jednadžba 2.27 izvodi operaciju na čvoru v_i koja kao ulazne parametre prima grupirane informacije $\mathbf{m}_i$ te reprezentaciju čvora $\mathbf{h}_i^l$ i daje ažuriranu reprezentaciju čvora $\mathbf{h}_i^{l+1}$.

Valja spomenuti da je izraz 2.26 samo jedan od mogućih načina agregacije informacija susjednih čvorova. Neke od najčešće korištenih metoda uz sumaciju jesu:

- Prosjek: izraz 2.26 dijeli se s brojem susjeda $\mathcal{N}(i)$, čime se zapravo uzima prosjek. To normalizira doprinos svakog susjeda, sprječavajući čvorove s mnogo susjeda da imaju neproporcionalno veliki utjecaj. Ovo je korisno ako želimo da svaki čvor ima jednak utjecaj bez obzira na broj susjeda, što je prikladno u situacijama gdje se stupnjevi čvorova značajno razlikuju.

- Maksimum: umjesto sumacije, uzima se samo maksimalna vrijednost $\mathbf{m}_{ij}$ u okolini atoma te ažurirana vrijednost postaje jednaka njoj. To omogućava modelu da se usredotoči na najistaknutije značajke susjeda te je korisna u scenarijima gdje najvažniji susjed određuje ishod.

- Minimum: sličan princip kao i maksimum samo što se uzima najmanja vrijednost skupa. Iako manje zastupljena od maksimuma, može biti korisna u slučajevima gdje je kritičan najmanji utjecaj ili najmanja vrijednost, primjerice u modeliranju najgoreg scenarija.

- Težinske sumacije: postoje razni oblici težinskih suma elemenata $\mathbf{m}_{ij}$ gdje težine c_{ij} mogu ovisiti ne samo o $\mathbf{m}_{ij}$, već i direktno o reprezentacijama čvorova $\mathbf{h}_i$ i $\mathbf{h}_j$. Ovakav pristup može biti koristan kada želimo naglasiti određene susjede više nego druge na temelju njihove relativne važnosti ili u drugom slučaju, kada model adaptivno treba naučiti te težine.

2.6 Nasumične Fourierove značajke

Autori članka [28] koristili su se nedavnim napretkom u modeliranju ponašanja dubokih mreža s pomoću regresije s kernelom i neuralnom tangentnom jezgrom (NTK od eng. Neural Tangent Kernel) [29] kako bi teorijski i eksperimentalno pokazali da standardni MLP-ovi nisu dobro prilagođeni za ove niskodimenzionalne zadatke u računalnom vidu i grafici zasnovanim na koordinatama. Konkretno, MLP-ovi imaju poteškoće u učenju funkcija visoke frekvencije, fenomen koji se u literaturi naziva spektralnom pristranošću (eng. spectral bias) [30, 31]. NTK teorija sugerira da je to zato što standardni MLP-ovi zasnovani na koordinatama odgovaraju kernelima s brzim padom frekvencija, što ih ograničava u predstavljanju visokofrekventnog sadržaja.

Nekoliko nedavnih radova [32,33] eksperimentalno je otkrilo da heuristično sinusoidno preslikavanje ulaznih koordinata (nazvano "pozicijsko kodiranje") omogućuje MLP-ovima da predstavljaju sadržaj viših frekvencija. Ulazne koordinate $\mathbf{v}$ bivaju preslikane u

$$\gamma(\mathbf{v}) = \left[a_1 \cos(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}),\, a_1 \sin(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}),\, ...,\, a_m \cos(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v}),\, a_m \sin(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v})\right]^{\mathrm{T}} \quad (2.28)$$

prije nego što su proslijeđene MLP-u. Pokazano je da ovo preslikavanje transformira NTK u stacionarnu (pomak-invarijantnu) jezgru i omogućuje podešavanje spektra NTK-a modificiranjem vektora frekvencije $\mathbf{b}_j$, čime se kontrolira raspon frekvencija koje odgovarajući MLP može naučiti. Jednostavna strategija postavljanja $a_j = 1$ i nasumičnog biranja $\mathbf{b}_j$ iz izotropne distribucije postiže dobre performanse, a skala (standardna devijacija) ove distribucije ima mnogo veći utjecaj na rezultate u usporedbi s njezinim specifičnim oblikom. U radu [28] su trenirani MLP-ovi s ovim Fourierovim preslikavanjem ulaza na nizu zadataka relevantnih za računalni vid i grafiku te je ovo preslikavanje dramatično poboljšalo performanse MLP-ova zasnovanih na koordinatama.

2.7 Finalni model i pristup korišten u radu

Hipoteza našeg je rada da kombinacijom dvaju setova podataka, različite točnosti i sličnih struktura, možemo dobiti bolji model nego korištenjem samo jednog od njih. Dva seta podataka koji će nam služiti jesu ANI-1x i ANI-1ccx, gdje je drugi očito onaj s preciznijim vrijednostima energije. Na raspolaganju imamo približno 4,6 milijuna parova molekula i njihovih energija iz ANI-1x te 500 tisuća iz ANI-1ccx. Za trening modela idealno bi bilo koristiti sve podatke na raspolaganju, no nažalost postoji ograničenje uzrokovano brzinom ovog procesa u kombinaciji s nedostatkom resursa. Iz tog razloga, uzeti su manji setovi podataka koji su po potrebi raspoređeni u skupove za trening, validaciju i testiranje. Više o ovom rasporedu bit će rečeno u sekciji rezultata i diskusije.

Budući da su ANI-1x i ANI-1ccx setovi podataka oboje sastavljeni od samo četiri elementa (H, C, N, O), za njihovo raspoznavanje neuralnoj bi mreži bilo dovoljno dati skalar iz nekog četveročlanog skupa, npr. $\{1, 6, 7, 8\}$. Iako bi ovo bilo dostatno, daleko je od najboljeg prikaza. Pristup koji je korišten u ovom radu jest prikazivanje svakog elementa s vektorom koji je ortogonalan na preostala tri. Kako su nam potrebna četiri međusobno ortogonalna vektora, njihova najmanja moguća dimenzija također je četiri. Ovakva transformacija u praksi se naziva ”one-hot-encoding”. Vektori koji su pridruženi elementima su

$$\mathrm{H} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathrm{C} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \mathrm{N} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad \mathrm{O} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{2.29}$$

Iako se na prvu ne čini kao značajna razlika, ovakav zapis omogućuje neuralnoj mreži jasnije raspoznavanje elemenata te povećava broj slobodnih parametara što dovodi do poboljšanja prediktivne moći mreže.

Korisno bi bilo da mreža ima mogućnost raspoznavanja podataka iz različitih setova. Umjesto da definiramo globalnu varijablu nad svakom molekulom, koja bi primjerice bila 1 za molekule iz ANI-1x i 0 za ANI-1ccx, ovom problemu možemo doskočiti tako da na vektore atoma manualno dodamo petu dimenziju koja će imati istu ulogu. Time prikaz elemenata molekula seta ANI-1x postaje

$$\mathrm{H} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad \mathrm{C} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad \mathrm{N} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \qquad \mathrm{O} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \tag{2.30}$$

a za ANI-1ccx

$$\mathrm{H} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathrm{C} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathrm{N} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \mathrm{O} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}. \tag{2.31}$$

Ovime smo kreirali ekvivalent globalne varijable, a da je ne moramo zasebno implementirati. Mreža sada razlikuje molekule tih dvaju setova. Valja prokomentirati ortogonalnost ANI-1x vektora elemenata koja je narušena dodavanjem pete dimenzije. Za razlikovanje atoma u molekuli njihovi pripadajući vektori ne moraju biti strogo ortogonalni, to vidimo u prvoj iteraciji rješenja gdje su elementi bili prikazani skalarima. Iako je ortogonalnost vektora u ovom slučaju poželjna, ona nije nužna te ne predstavlja problem pri raspoznavanju atoma.

Želimo li poboljšati rezultate očito neće biti dovoljno koristiti se običnim GNN-om, potreban nam je neki drugačiji pristup. Iz tog razloga, korišten je EGNN u kombinaciji s RFF-om. Kao što je već spomenuto u sekciji 2.5, korištenje EGNN pristupa omogućuje interakciju čvorova koji nisu direktno međusobno povezani. Imamo slobodu odabira koliko će duboka biti ova interakcija, tj. efektivno možemo na to gledati kao pitanje: koji će red susjeda biti uključen u interakciju? To reguliramo odabirom broja EGNN slojeva koji će se odviti za svaku epohu neuralne mreže. U prvom ciklusu svaki atom biva ažuriran informacijama svojih susjeda. U drugom ciklusu kada atom prima informaciju susjeda, ona sada sadrži i informacije njegovih susjeda pa na to možemo gledati kao da početni atom interagira sa svojim drugim susjedom itd. Naravno, povećanjem broja tih slojeva, poboljšavamo prediktivnu moć mreže i idealno bi bilo kada bi svaki atom mogao biti ažuriran informacijama svih ostalih, no to dolazi s velikom računskom cijenom. Empirički je pokazano da je postignut dobar

omjer prediktivne moći mreže i njezine brzine za 3 sloja. Kako koordinate atoma bivaju ažurirane svakim slojem, mreži nije dovoljno dati informacije o molekuli u obliku grafa već nas zanimaju egzaktne pozicije atoma. Korištena metoda agregacije informacija čvorova je sumacija. Ovaj izbor nam najviše odgovara jer želimo imati kumulativan utjecaj svih susjeda.

Za prikaz udaljenosti među atomima koristimo se RFF-om. Ovaj pristup pruža skalabilan i fleksibilan pristup uključivanju nelinearnih odnosa između udaljenosti čvorova. Za razliku od izravnog pretvaranja udaljenosti u Gaussove značajke koje nude fiksnu i ograničenu transformaciju, RFF omogućuje projekciju ulaznih udaljenosti u višedimenzionalni prostor. Ovo poboljšava izražajnost mreže, omogućavajući joj modeliranje složenijih interakcija između čvorova. Osim toga, RFF podržava bolju generalizaciju aproksimacijom šireg spektra funkcija, što ga čini superiornim u scenarijima gdje je ključno uhvatiti složene uzorke u podatcima. Skalabilnost RFF-a, zajedno s mogućnošću kontrole dimenzionalnosti, čini ga vrijednim alatom za poboljšanje performansi grafskih neuronskih mreža. Ovdje imamo slobodu odabira broja Fourierovih parova, tj. dimenzije vektora koji će prikazivati udaljenost između dva atoma te faktora skaliranja (a_j iz sekcije 2.6). U ovom radu vrijednost skalirajućeg faktora postavljena je na 1, a dimenzija vektora udaljenosti je 64.

Valja još spomenuti korištenu metodu "sažimanja" (eng. pooling). Nakon što informacije prođu kroz EGNN slojeve, svaka značajka na razini čvora u grafu (što predstavlja atom u molekuli) sadrži informacije susjednih čvorova. Međutim, za mnoge zadatke, uključujući predviđanje svojstava molekula poput energije, cilj je dobiti jedan izlaz za cijeli graf (molekulu), a ne zasebne izlaze za svaki čvor. Proces kojim se odvija to združivanje informacija naziva se sažimanje. Postoji nekoliko metoda sažimanja i u suštini sve rade na istom principu kao i agregacija informacija u slojevima EGNN-a iz sekcije 2.6. Moguće je uzeti sumu vektora čvorova, njihov prosjek, samo maksimalnu ili minimalnu vrijednost svih čvorova, ili neku vrstu težinske sume. Izbor u ovom radu bila je obična sumacija čiji egzaktan izraz glasi:

$$\mathbf{h}_g = \sum_{i \in \mathcal{V}} \mathbf{h}_i \tag{2.32}$$

gdje se $\mathbf{h}_g$ odnosi na graf, $\mathbf{h}_i$ na pojedini čvor, a $\mathcal{V}$ na skup svih čvorova grafa. Budući da je energija sustava po sebi aditivno svojstvo, tj. ukupna energija zbroj

je pojedinačnih doprinosa svih atoma, sumarno sažimanje najbolje odgovara našim potrebama. Još je jedno ključno svojstvo ovog pristupa invarijantnost na redoslijed čvorova u grafu. To znači da će zbroj, bez obzira na to kako su čvorovi indeksirani, uvijek biti isti te time i predikcija energije za taj sustav.

Također, kao korak predobrade podataka, oduzeti su doprinosi ukupnoj energiji od pojedinih atoma. Ovaj korak pomaže u izoliranju energija vezanja od inherentnih atomskih energija, što osigurava usporedivost između različitih skupova za treniranje. Time se također omogućuje preciznija analiza međuatomskih interakcija, bez utjecaja osnovnih energetskih doprinosa pojedinih atoma. Python skripta za ovaj proces te svi ostali kodovi potrebni za treniranje ove neuralne mreže nalaze se u dodatku.

Takav model bi se, osim na molekule, mogao primijeniti i na kristale. Primjerice, kombinacijom jednog od ANI setova s molekularnim kristalima (sastavljenim samo od H, C, N, O) možemo preciznije predvidjeti njihova svojstva. EGNN bi nam ponovo osigurao interakciju udaljenih atoma, a svojstvo periodičnih funkcija RFF-a bilo bi još povoljnije zbog periodičkih svojstava kristala.

3 Rezultati i diskusija

3.1 Raspodjela podataka

Korišten je model opisan u sekciji 2.7 koji je baziran na kombinaciji EGNN-a s RFF-om uz neke dodatne modifikacije koje mu omogućuju primanje podataka iz različitih setova. Ovakav pristup donosi dodatnu kompleksnost mreži u nadi da će time ona mogući prepoznati složene detalje međuatomske interakcije te poboljšati postojeći model. Kao što je prije spomenuto, korištena je kombinacija molekula setova ANI-1x i ANI-1ccx te je sveukupno pokrenuto 8 različitih treninga s rasporedom podataka prikazanim u tablici 3.1.

Modeli	**Trening**		**Validacija**		**Testiranje**	
	ANI-1x	**ANI-1ccx**	**ANI-1x**	**ANI-1ccx**	**ANI-1x**	**ANI-1ccx**
ANI-1x	160k	0	20k	0	0	20k
ANI-1ccx	0	20k	0	20k	0	20k
20k/20k	20k	20k	10k	20k	0	20k
80k/20k	80k	20k	10k	20k	0	20k
160k/20k	160k	20k	20k	20k	0	20k
20k/60k	20k	60k	10k	20k	0	20k
80k/60k	80k	60k	10k	20k	0	20k
160k/60k	160k	60k	20k	20k	0	20k

Tablica 3.1: Raspodjela podataka po modelima.

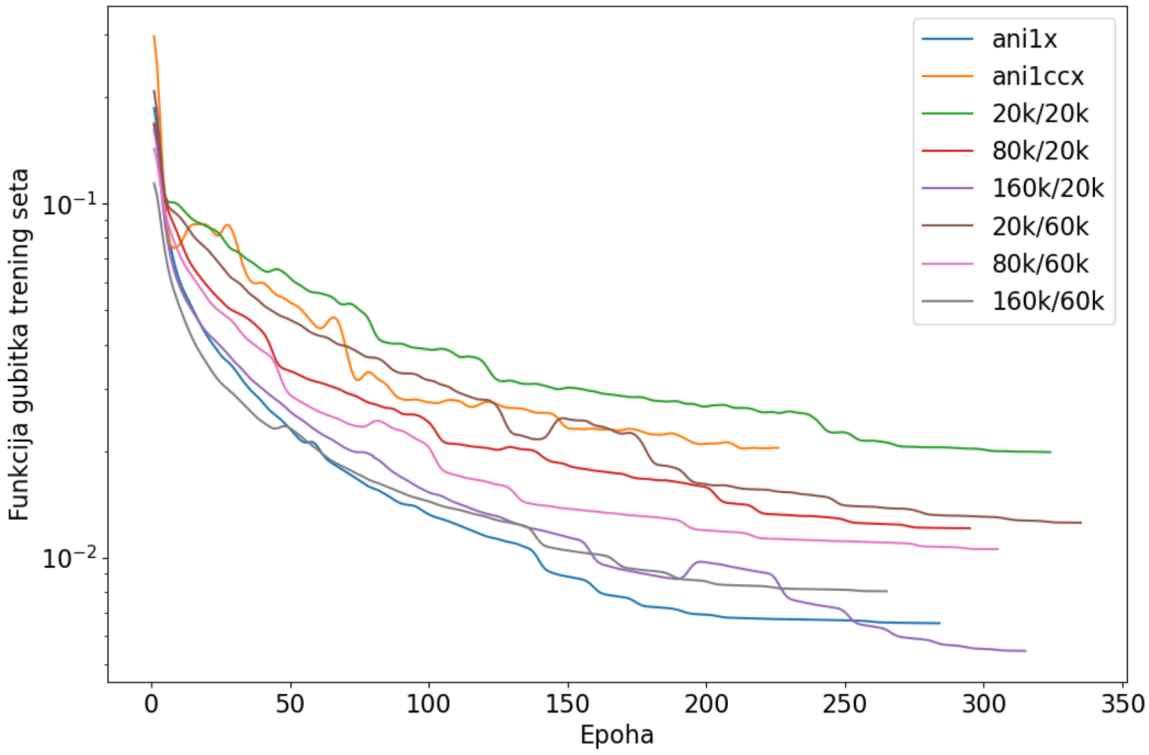
Ideja je bila usporediti preciznost modela s obzirom na omjer podataka iz ta dva seta. Za te potrebe pokrenute su dvije grupe modela u kojima je količina ANI-1ccx podataka držana konstantna, a količina ANI-1x bila je promjenljiva. Također, usporedbe radi pokrenuta su i dva modela koja ne sadrže kombinirane podatke; ANI-1x i ANI-1ccx sadrže samo istoimene podatke u setovima za trening i validaciju. Za sve modele, test set se sastoji samo od ANI-1ccx podataka jer nas zanima predikcija energije samo nad njima.

Kombinacija podataka u setu validacije ima najmanji utjecaj na performanse jer je on korišten samo za regulaciju stope učenja (eng. learning rate). Ukoliko se SKP nad tim setom ne smanji u 10 uzastopnih epoha, stopa učenja biva umanjena za faktor 2. Početna vrijednost postavljena je na 10^{-4}, a treniranje modela završava kada njena vrijednost padne ispod 10^{-7}.

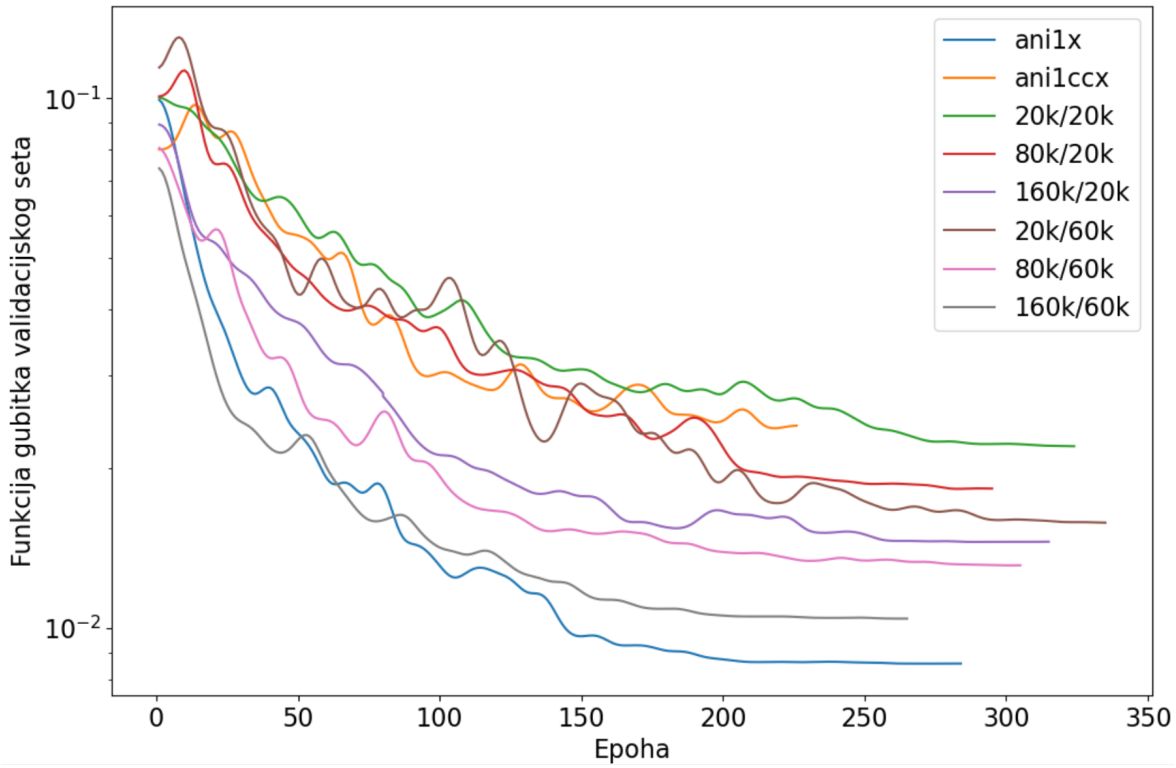
Broj podataka koje smo izabrali iz svakog seta je u potpunosti proizvoljan, no količina ANI-1x podataka veća je ili jednaka količini ANI-1ccx za većinu modela. Ovo i je realan razmjer ako se uzmu u obzir svi dostupni podatci. Također, želimo li primijeniti model na neki drugi sustav, gotovo će uvijek podatci kojima želimo prilagoditi model biti u manjini. Finalne količine podataka uzete za ovaj rad bazirane su na odnosu brzine treniranja modela sa zadovoljavajućom količinom podataka za generalizaciju problema.

3.2 Rezultati

Promotrimo ovisnost funkcije gubitka (eng. loss function) o broju epohe za trening i validaciju (slike 3.1 i 3.2). Odabrana funkcija gubitka bio je korijen prethodno definiranog SKP-a. Za prikaz funkcija gubitka korišteno je Gaussovo zaglađivanje s pomoću funkcije gaussian_filter1d iz paketa scipy.ndimage [34]. Gaussovo zaglađivanje bazirano je na uprosječivanju obližnjih točaka, pri čemu one teže Gaussovoj distribuciji. Za funkciju gubitka trening seta korištena je vrijednost sigma $= 2$, a za funkciju gubitka validacijskog seta sigma $= 4$. Radi bolje preglednosti željeli smo izbjeći nazubljenost funkcija gubitaka do koje dolazi zbog oscilacija.



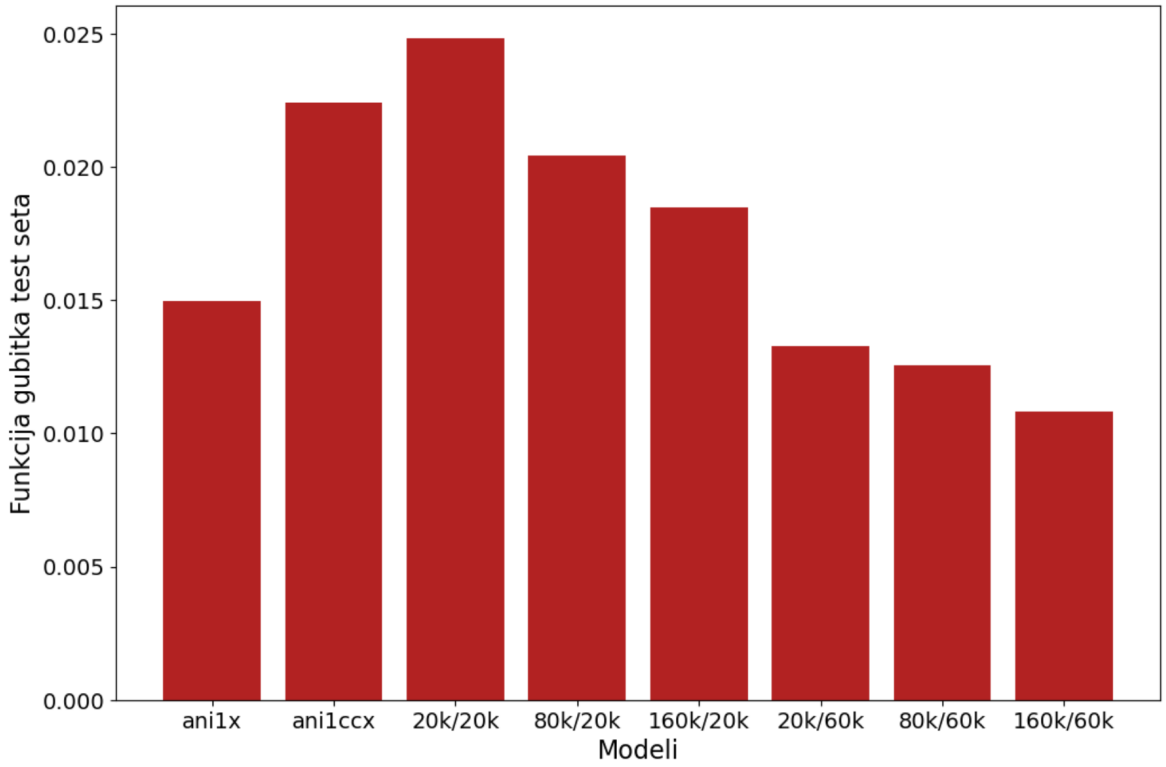
Slika 3.1: Ovisnost funkcije gubitka na trening setu o epohi za različite kombinacije podataka.



Slika 3.2: Ovisnost funkcije gubitka na validacijskom setu o epohi za različite kombinacije podataka.

Slike 3.1 i 3.2 prikazuju ovisnosti funkcije gubitaka za trening i validaciju za različite modele. Očekivano, ako držimo količinu ANI-1ccx podataka fiksnom, vrijednosti funkcije gubitka na trening setu i setu za validaciju za kombinirane podatke ponašaju se obrnuto proporcionalno s količinom ANI-1x podataka. Ono što nas primarno zanima jest usporedba modela nad setom za testiranje. Budući da su svi test setovi identični to će biti direktan pokazatelj preciznosti modela. Ova usporedba prikazana je na slici 3.3.

Ovdje jasno vidimo odnos ”kvalitete” modela. Dodavanje malog broja ANI-1x podataka učinilo je model neznatno lošijim. Svako sljedeće povećanje trening seta rezultiralo je poboljšanjem modela, te modeli 80k/20k i 160k/20k pokazuju bolje performanse od samog ANI-1ccx. Povećanjem količine ANI-1ccx podataka u setu za učenje mreže, no smanjivanjem količine ANI-1x podataka također rezultira poboljšanjem modela. Naravno, ovaj rezultat ovisi o faktorima za koje smo skalirali te setove. U slučaju u kojem bi povećanje bilo minimalno, a smanjenje drastično, rezultiralo bi lošijim modelom. Pri većoj količini ANI-1ccx podataka u setu za učenje, također vidimo isti trend smanjenja funkcije gubitka povećavanjem količine ANI-1x



Slika 3.3: Vrijednost funkcije gubitka na test setu za različite kombinacije podataka.

podataka. Zaključujemo da kombinacija setova treniranja uistinu poboljšava model.

Valjalo bi još prokomentirati model treniran samo na ANI-1x podatcima. Očekivano, vrijednosti funkcije gubitka nad setovima treninga i validacije među nižima su kad ih usporedimo s drugim modelima. Razlog tomu je taj što model sadrži dovoljnu količinu samo jednog tipa podataka da mreža uspije ”naučiti” svojstva molekula nad kojima se funkcija izvrjednjuje.

Na prvu neočekivan je rezultat dobiven za funkciju gubitka na test setu modela koji je treniran samo na ANI-1x podatcima kada ga usporedimo s ostalima. Postavlja se pitanje zašto dodavanje manje količina ANI-1ccx podataka čini model lošijim iako on biva testiran upravo nad njima. Objašnjenje leži u sastavu ANI-1ccx seta. Prisjetimo se da ANI-1x, pa tako i ANI-1ccx, sadrže pažljivo probrane molekule od kojih svaka nosi neku novu informaciju korisnu za treniranje mreže. Uzmemo li samo dio tih molekula za treniranje mreže, on može imati negativan efekt na finalne performanse upravo zato što će se set za testiranje značajno razlikovati od njega. Daljnjim povećanjem količine podataka ANI-1ccx seta ovaj se efekt smanjuje jer mreža ima pristup većoj raznolikosti molekula, što je i vidljivo u performansama modela sa 60 000 podataka ovog seta.

Također, vrijednosti funkcija gubitka na test setu u zadovoljavajućim su granicama. Obično se greške neuralnih mreža kreću oko vrijednosti od 1 meV/atom. Greške ovih modela kreću se od 25 do 12 meV po molekuli. Uzmemo li u obzir da se prosječna molekula ovih setova sastoji od nešto manje od 20 atoma, dobivene pogreške kreću se oko ciljane vrijednosti od jednog meV po atomu.

3.3 Diskusija

Preostaje još diskutirati o nekoliko detalja u vezi s ovim procesom i njegovom generalizacijom. Promotrimo odnos vremena treninga s obzirom na količinu podataka. Veličine trening setova u omjeru su $2:5:9$ (za kombinirane setove s 20k ANI1-ccx), dok je odnos potrebnog vremena po epohi otprilike identičan tome. Drugim riječima, vrijeme treniranja modela skalira linearno s količinom podataka. Ista ovisnost vremena i količine podataka zapažena je kod setova sa 60 tisuća ANI1-ccx podataka u trening setu.

Također, postoji problem odabira test podataka. Iako su svi modeli na kraju testirani nad istim podskupom ANI-1ccx seta, ovo nije savršen prikaz njihovih performansi. Kao što je spomenuto u sekciji 2.4, molekule ovog seta raznovrsne su i pažljivo probrane kako bi učinile originalni model što boljim uz što manje podataka. S time na umu, postoji mogućnost da bi odnosi naših modela bili nešto drugačiji u slučaju da je bio izabran neki drugi podskup. Iako to ne bi dovelo do drastičnih razlika, važno ih je biti svjestan prije donošenja zaključaka oko generalizacije procesa.

Ovakav model ima mnogo hiperparametara koji bi mogli biti korigirani u korist njegove performanse. Iako je model dao zadovoljavajuće rezultate s trenutnim postavkama, ne možemo biti sigurni da su one najbolji izbor. Primjerice, broj EGNN slojeva te dimenzija RFF vektora, direktno definiraju strukturu mreže. Također, postoji pitanje optimizacije količine ANI-1x podataka. Naravno, to ovisi o tome ciljamo li na što bolje performanse modela ili na brzinu treniranja. Budući da nas većinom zanima balans tih dviju veličina, postoji neka granična količina podataka nakon koje model ne daje značajno bolje rezultate, a postaje računski skuplji.

Naposljetku, komentirajmo još aplikacije ovog modela na druge sustave. Sada kada smo pokazali da je pristup valjan, pitanje je gdje bismo ga još mogli primijeniti. Drugim riječima, koliko različiti mogu biti setovi podataka koje kombiniramo, a da

njihova kombinacija i dalje rezultira poboljšanjem modela. Kao što je prethodno spomenuto, sustav na koji bi ovaj model mogao biti primjenljiv jesu molekularni kristali. Zbog njihovih brojnih konformacija te međuatomskih interakcija (Van der Waals-ove i vodikove veze) koje je teško modelirati, količina je dostupnih podataka za ove sustave ograničena. Kako broj potrebnih primjera za treniranje raste s njihovom kompleksnošću, teško je napraviti precizan model baziran samo na molekularnim kristalima. Rješenje ovog problema mogla bi biti kombinacija tih kristala s ANI-1 setom.

4 Zaključak

U ovome smo se radu koristili neuralnim mrežama baziranim na kombinaciji EGNN-a s RFF-om. Ekvivarijantni pristup omogućuje interakciju čvorova (koji predstavljaju atome) koji nisu direktno međusobno povezani. Za potrebe našeg modela, dubina interakcije bila je postavljena do trećeg susjeda. Izbor prikaza udaljenosti atoma, korištenjem Fourierovih umjesto Gaussovih članova, omogućuje nam stabilan i fleksibilan pristup uključivanju nelinearnih odnosa među njima. Korištena su dva seta podataka, ANI-1x i ANI-1ccx, koji imaju različite razine točnosti. Na raspolaganju imamo otprilike četiri puta više ANI-1x podataka, koji ujedno imaju i manju preciznost vrijednosti energija.

Cilj rada bio je pokazati da kombinacijom ovih dvaju setova dobiveni model ima bolje performanse od onog treniranog samo na setu s preciznijim podatcima (ANI-1ccx). Za te potrebe trenirano je osam modela: jedan samo na ANI-1x podatcima, jedan samo na ANI-1ccx podatcima, a preostalih je šest imalo različite količine dodanih ANI-1x podataka setovima za trening i validaciju. Svi setovi za testiranje modela identični su te se sastoje samo od ANI-1ccx molekula.

Prije početka treninga svim molekulama oduzeti su doprinosi ukupnoj energiji od pojedinih atoma. Ovo je napravljeno kako bi se izolirala energija vezanja od inherentnih atomskih energija, što osigurava usporedivost između različitih skupova za treniranje. Također, mreži je dana mogućnost razlikovanja molekula na temelju skupa podataka iz kojeg proizlaze.

Zasebno gledajući, svi modeli su dali zadovoljavajuće rezultate. Vrijednosti funkcije gubitka na test setu kreću se od 25 do 12 meV po molekuli. Budući da prosječna molekula ovih setova ima 20-ak atoma, dobivena greška iznosi oko 1 meV po atomu, što je u skladu s današnjim modelima. Trenirane su dvije grupe po tri modela u kojima je količina ANI-1ccx podataka u trening setu držana konstantnom. Prva grupa sadržavala je 20 tisuća ANI-1ccx primjera, a omjeri s obzirom na ANI-1x bili su 1:1, 1:4 i 1:8. Druga grupa bila je sačinjena od 60 tisuća ANI-1ccx molekula, a omjeri s obzirom na ANI-1x bili su 3:1, 3:4 i 3:8. Kod svih modela s kombinacijom podataka, povećanjem broja ANI-1x podataka u trening setu poboljšavala se i prediktivna moć mreže. Unatoč tome, prva kombinacija dala je lošije rezultate od modela treniranog samo na ANI-1ccx, no sve ostale su ga nadmašile.

Iako postoji još prostora za napredak, što bi zahtijevalo dodatna testiranja, pokazano je da kombiniranje podataka različite točnosti uistinu poboljšava model. Postavlja se pitanje gdje bi još ovakav pristup bio koristan. Jedna primamljiva opcija bili bi molekularni kristali. Njihova kompleksnost čini ih izazovnima za modeliranje te je količina dostupnih podataka za ove sustave ograničena. Njihovom kombinacijom s ANI-1 setovima potencijalno bi se riješio problem manjka podataka za treniranje neuralnih mreža samo na molekularnim kristalima.

Dodatak

Poveznica za GitHub repozitorij u kojem se nalaze sve skripte korištene u izradi ovog rada. Repozitorij sadrži README.md dokument koji ima više detalja o pojedinim skriptama.

`https://github.com/FranK1308/Diplomski`

Literatura

[1] Dronskowski, R. Computational Chemistry of Solid State Materials. 1st ed. : Wiley-VCH, 2005.

[2] Vink, R. L. C.; Barkema, G. T.; Stijnman, M. A.; Bisseling, R. H. Physics Review B : porpose-led publishing, 2001.

[3] Hobday, S.; Smith, R.; Belbruno, J. Applications of neural networks to fitting interatomic potential functions, Vol 7, Number 3 : porpose-led publishing, 1999.

[4] Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces // Physical review letters 98, 2007.

[5] Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. // Physical Chemistry Chemical Physics. Vol. 13, 40 (2011), str. 17930–17955.

[6] Behler, J.; Martoňák, R.; Donadio, D. et al. Pressure-induced phase transiti ons in silicon studied by neural network-based metadynamics simulations. // physica status solidi (b). Vol. 245, 12 (2008), str. 2618–2629.

[7] Behler, J.; Martoňák, R.; Donadio, D. et al. Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. // Physical review letters. Vol. 100, 18 (2008), str. 185501.

[8] Smith, J. S.; Isayev, O.; Roitberg, A. E. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. // Chemical science. Vol. 8, 4 (2017), str. 3192–3203.

[9] Behler, J. Perspective: Machine learning potentials for atomistic simulations. // The Journal of chemical physics. Vol. 145, 17 (2016), str. 170901.

[10] Karlik, B.; Olgac, A. V. Performance analysis of various activation functions in generalized mlp architectures of neural networks. // International Journal of Artificial Intelligence and Expert Systems. Vol. 1, 4 (2011), str. 111–122.

[11] Fink, T.; Raymond, J. L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stere-

oisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery // Journal of Chemical Information and Modeling. Vol 747 2(2007), str. 342–353.

[12] Fink, T.; Bruggesser, H.; Raymond, J. L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons // Angewandte Chemie International Edition. Vol 44, 10(2005), str. 1504-1508.

[13] Chai, J.;Head-Gordon, M. The Journal of Systematic optimization of long-range corrected hybrid density functionals // The Journal of Chemical Physics. Vol 128, 8(2008)

[14] Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules // The Journal of Chemical Physics. Vol 54, 2(1971)

[15] Smith, J. S.; Nebgen, B.; Lubbers, N. et al. Less is more: Sampling chemical space with active learning. // The Journal of chemical physics. Vol. 148, 24(2018), str. 241733.

[16] Jupp, S.; Malone, J.; Bolleman, J.; Brandizi, M.; Davies, M.; Garcia, L.; Gaulton, A.; Gehant, S.; Laibe, C.; Redaschi, N.; Wimalaratne, S. M.; Martin, M.; Le Novre, N.; Parkinson, H.; Birney, E.; Jenkinson, A. M. The EBI RDF platform: linked open data for the life sciences // Bioinformatics. Vol. 30, 9(2014), str. 1338-1339.

[17] Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update // Nucleic Acids Research. Vol 42, D1(2014), str. D1083–D1090.

[18] Davies, M.; Nowotka, M.; Papadatos, G.; Atkinson, F.; van Westen, G.; Dedman, N.; Ochoa, R.; Overington, J. MyChEMBL: A Virtual Platform for Distributing Cheminformatics Tools and Open Data // Challenges 5, 334(2014).

[19] Seung, H. S.; Opper, M.; Sompolinsky, H. Query by committee. Proceedings of the fifth annual workshop on Computational learning theory 1992 str. 287–294.

[20] Purvis, G. D.; Bartlett, R. J. A full coupled-cluster singles and doubles model: the inclusion of disconnected triples // The Journal of Chemical Phyics. 76(1982), str. 1910–1918.

[21] Bartlett, R. J.; Musiał, M. Coupled-cluster theory in quantum chemistry // Reviews of Modern Physics 79(2007), str. 291–352.

[22] Daniel Crawford, T.; F. Schaefer, H. III An introduction to coupled cluster theory for computational chemists // Reviews in Computational Chemistry 14(2007), str. 33–136.

[23] Hobza, P.; Šponer, J. Toward true DNA base-stacking energies: MP2, CCSD(T), and complete basis set calculations // Journal of American Chemical Society 124(2002), str. 11802–11808.

[24] Feller, D.; Peterson, K. A.; Crawford, T. D. Sources of error in electronic structure calculations on small chemical systems // The Journal of Chemical Phyics 124(2006).

[25] Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 data set: more accurate interaction energies and angular-displaced nonequilibrium geometries // Journal of Chemical Theory and Computation 7, 11(2001), str. 3466–3470.

[26] Smith, J.S.; Nebgen, B.T.; Zubatyuk, R. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning // Nature Communications Vol. 10, 2903(2019).

[27] Satorras, V. G.; Hoogeboom, E.; Welling, M. Proceedings of the 38th International Conference on Machine Learning : E(n) Equivariant Graph Neural Networks // Proceedings of Machine Learning Research

[28] Tancik, M.; Srinivasan, P. P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J. T.; Ng, R. Advances in Neural Information Processing Systems 33 : Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains // Conference on Neural Information Processing Systems (2020)

[29] Jacot, A.; Gabriel, F.; Hongler, C. Neural Tangent Kernel: Convergence and generalization in neural networks // Advances in Neural Information Processing Systems 31, 2018.

[30] Basri, R.; Galun, M.; Geifman, A.; Jacobs, D.; Kasten, Y.; Kritchman, S. Proceedings of the 37th International Conference on Machine Learning: Frequency bias in neural networks for input of non-uniform density // Proceedings of Machine Learning Research (2020)

[31] Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F. A.; Bengio, Y.; Courville, A. On the spectral bias of neural networks // International Conference on Machine Learning, 2019.

[32] Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis // Communications of the Association for Computing Machinery Vol. 65, Issue 1 (2021)

[33] Zhong, E. D.; Bepler, T.; Davis, J. H.; Berger, B. Reconstructing continuous distributions of 3D protein structure from cryo-EM images // International Conference on Learning Representations (2020)

[34] SciPy library: Multidimensional image processing (ndimage) `https://docs.scipy.org/doc/scipy/tutorial/ndimage.html`